EXAMINING THE TEMPORAL DYNAMICS OF HUMAN REPRESENTATIONAL ALIGNMENT BY PROBING MODEL LEARNING WINDOWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Critical periods in biological development are temporal windows when experiences leave lasting effects on a system. Timing of critical periods differs based on the function being learned, with different sensory, cognitive, and linguistic systems following distinct schedules. In deep neural networks, critical periods have been identified primarily through task accuracy, collapsing this diversity into a single curve. We ask whether critical periods differ when measured through representational alignment with the human visual cortex. We train ResNet, ConvNeXt, and ViT under supervised ImageNet classification, ViT under contrastive language-image pretraining (CLIP), and pretrained CLIP-ViT under a fine-tuning task, injecting transient perturbations to inputs and targets at different training stages. We measure the effects at three levels: immediate change in internal representations (via similarity to human behaviorally-derived embeddings and fMRI activations), the trajectory representations follow as normal training resumes, and generalizability of learned embeddings via linear probes on out-of-distribution tasks. Early input-side perturbations cause the most disruption to representational structure, and models that recover task performance also recover human alignment. In supervised classifiers, human alignment and task performance recover from all perturbations. In CLIP, early and mid perturbations permanently damage task accuracy, yet after mid-training perturbations, brain alignment recovers in every case and behavioral alignment in two of three, showing the two objectives can dissociate. Late input-side perturbations have a different effect on ImageNet classifiers: representations of within-category objects are pulled closer together, distinguishing between-category boundaries and transiently improving human alignment. Critical periods in neural networks depend on what is measured, not just when. If late perturbations push representations toward human-like organization with little harm to performance, targeted interventions could steer representational alignment.

1 INTRODUCTION

Learning in both biological and artificial systems unfolds through temporally sensitive phases. In biological development, *critical periods* are windows of heightened plasticity in which experience has disproportionate and sometimes irreversible effects on cortical organization (Hubel & Wiesel, 1970; Hensch, 2004). Neural networks also have sensitive early training phases, in which noise, data removal, or regularization can permanently change the model (Achille et al., 2019; Kleinman et al., 2024; Fukase et al., 2025; Lee et al., 2026), but these effects have been measured mainly through task loss and accuracy. How do such interventions affect how human-like the representations become? Human-like representations can make models more robust and generalizable (Muttenthaler et al., 2023; Sucholutsky & Griffiths, 2023; Muttenthaler et al., 2025), so knowing when human-likeness is most sensitive to intervention, and whether that differs from task performance, matters for training interventions that serve both objectives.

Following Achille et al. (2019) and Lee et al. (2026), we apply image-side (Gaussian image noise, uniform images) and target-side (shuffled labels, random targets) perturbations early, mid, or late in training. Our contributions are:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

- **A perturbation framework for alignment.** We perturb one epoch at controlled points in the training of three ImageNet classifiers (RN50, ViT-B/16, ConvNeXt-B), a CLIP ViT-B/32 trained from scratch, and a fine-tuned CLIP-HBA-Behavior, and track behavioral and neural alignment, accuracy, and linear-probe transfer (Section 3).
- **Accuracy and alignment dissociate.** A run can lose nearly all of its accuracy yet keep its alignment, or the reverse. Supervised classifiers recover both, usually within an epoch, whereas CLIP ViT-B/32 keeps lasting accuracy deficits after early and mid perturbations, yet recovers its alignment after the mid perturbations (Sections 4.2 and 4.3).
- **A class-versus-category account.** Accuracy depends more on class-level and alignment more on category-level structure. Preserving the arrangement of categories and the separation of classes within them predicts retained accuracy, and keeping categories separable predicts retained alignment (Section 4.2).
- **Late perturbations transiently raise alignment.** In ViT-B/16 and ConvNeXt-B, whose baselines trade category-level for class-level structure late in training, late image-side perturbations raise alignment with little harm to accuracy until normal training resumes (Section 4.1).

2 Related work

2.1 Critical periods in biological and artificial systems

Critical periods recur in biological development across systems and species. One classic example is ocular dominance in kittens, established by Hubel & Wiesel (1970). They found that briefly closing one eye during a defined postnatal window permanently reshapes cortical responses, while the same monocular deprivation in adulthood leaves learned cortical responses intact. The same phenomenon appears in auditory-visual spatial mapping in the barn owl, where Brainard & Knudsen (1998) found that barn owls calibrate their auditory spacial map to align with vision only during a juvenile window. In a similar vein, the white-crowned sparrow must hear a song within the first 10–50 days of life in order to develop normal song (Marler, 1970). With regard to language, Werker & Tees (1984) showed that infants can discriminate non-native speech contrasts, but lose this ability in the first year of life. These are only a few examples of the myriad critical periods that occur across modalities, systems, functions, species, and individuals. This consistency suggests critical periods are not an idiosyncrasy of any one circuit, but a general feature of how experience-dependent systems learn (see Knudsen, 2004).

While biochemical processes are partially responsible for modulating biological critical periods, the system's architecture also plays a role in how critical periods form over development (see Hensch, 2004). Indeed, Achille et al. (2019) showed that critical periods occur even in artificial neural networks, where no such biochemistry exists. They show that experience-dependent learning involves an early phase when maximum structure is absorbed from the environment, followed by a consolidation phase where the network distills and commits that information. If the early phase involves learning on a flawed data distribution, the network cannot recover the same solution in the loss landscape as would have been discovered if the system had been trained on the true distribution. This consolidation process helps explain why deep neural networks have difficulty acquiring new information later in training, yet consolidation also plays a crucial role in forming *sufficient*, *minimal*, and *invariant* representations (Achille & Soatto, 2018). This same late-training pruning can also contribute to model *overfitting* if practitioners are not careful.

2.2 Human representational alignment and vision models

Alignment is worth measuring separately from accuracy, as the two can diverge. In training more than 600 models, Gokce & Schrimpf (2025) found that behavioral alignment generally follows accuracy, while neural alignment saturates. Perceptual similarity in ResNets peaks within a few epochs and then falls as accuracy rises (Kumar et al., 2022); and more accurate ImageNet models tend to be worse models of human visual strategies and of primate IT (Fel et al., 2022; Linsley et al., 2023). Alignment may also be a valuable objective. Models trained on human label distributions (Peterson et al., 2019), human-aligned pretrained models (Sucholutsky & Griffiths, 2023), and models fine-tuned toward human similarity judgments (Muttenthaler et al., 2025) or macaque IT (Dapello et al.,

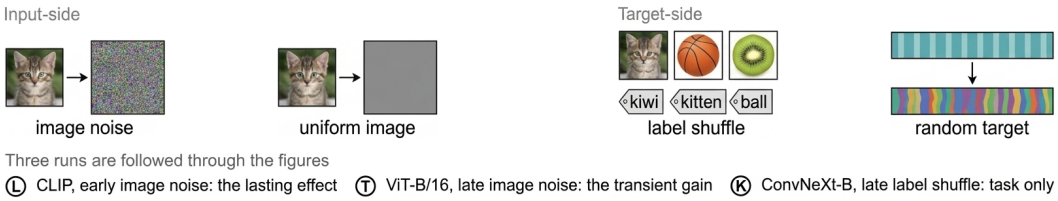

Figure 1: **Each perturbation corrupts a single epoch of otherwise unchanged training.** Image noise and uniform images are input-side; label shuffle and random target (CLIP-HBA-Behavior only) are target-side. Each perturbed run is compared with its seed-matched baseline on ImageNet top-1, linear-probe transfer, behavioral alignment (SPoSE), and brain alignment (fMRI).

2023) are more robust or generalize better, although fine-tuning on mid-level NIGHTS judgments (Fu et al., 2023) hurts classification and fine-tuning on THINGS triplets worsens transfer (Sundaram et al., 2024). These studies either track alignment over unperturbed training or impose it through the objective or after pretraining. Several show when alignment emerges, but, to date, no study perturbs training at controlled times to ask when human-like structure is most sensitive.

3 Methods

3.1 Models and datasets

We train three ImageNet-1k classifiers (Deng et al., 2009; Russakovsky et al., 2015) from scratch for 90 epochs with cross-entropy: ResNet50 (RN50; He et al., 2016), ViT-B/16 (Dosovitskiy et al., 2021), and the modernized convolutional ConvNeXt-B (Liu et al., 2022). We train CLIP-ViT-B/32 (Radford et al., 2021) contrastively for 32 epochs on YFCC15M, a 15-million-pair English subset of YFCC100M (Thomee et al., 2016), and fine-tune CLIP-HBA-Behavior (Zhao et al., 2025), a parameter-efficient CLIP-ViT-L/14 trained to predict human SPoSE embeddings (Appendix B.2). Hyperparameters are in Appendix A.

Behavioral alignment is evaluated on THINGS (Hebart et al., 2019): 1,854 object concepts, each with an image and a sparse positive similarity embedding (SPoSE) whose 66 dimensions are interpretable features derived from human judgments, such as “metallic” or “animal-related” (Appendix B.1).

3.2 Behavioral and brain alignment metrics

We measure alignment with representational similarity analysis (RSA; Kriegeskorte et al., 2008), which compares representational dissimilarity matrices (RDMs) without requiring a shared coordinate system, and repeat every behavioral-alignment analysis with centered kernel alignment (CKA; Kornblith et al., 2019a) in Appendix G; the two have complementary invariances (Appendix D).

Behavioral alignment, computed at every epoch, is the Spearman ρ between the model’s RDM and the human RDM over the 1,854 THINGS images, where each entry is $1-$ Pearson r between two images’ representations: the model’s penultimate-layer features (for CLIP-ViT-B/32, its 768-dimensional class-token features before the image–text projection) and the objects’ 66-dimensional SPoSE embeddings. Images are preprocessed as in each model’s training pipeline.

Brain alignment compares models with subject-averaged region-of-interest (ROI) RDMs of fMRI responses: 27 ROIs over 720 concepts from THINGS-fMRI and, as an out-of-distribution replication, 14 retinotopic and object-selective ROIs from the Natural Object Dataset (NOD, Gong et al., 2023). Because early visual areas may be best predicted by early layers and high-level areas by late ones, we use *layer-reweighted RSA* (LR-RSA), adapted from feature-reweighted RSA (Kaniuth & Hebart, 2022): a non-negative ridge regression reweights RDMs from every residual block to predict each ROI’s RDM (Appendix D). Brain alignment is the in-sample Spearman ρ between each ROI’s RDM and its fitted combination, averaged over the 27 THINGS-fMRI ROIs.

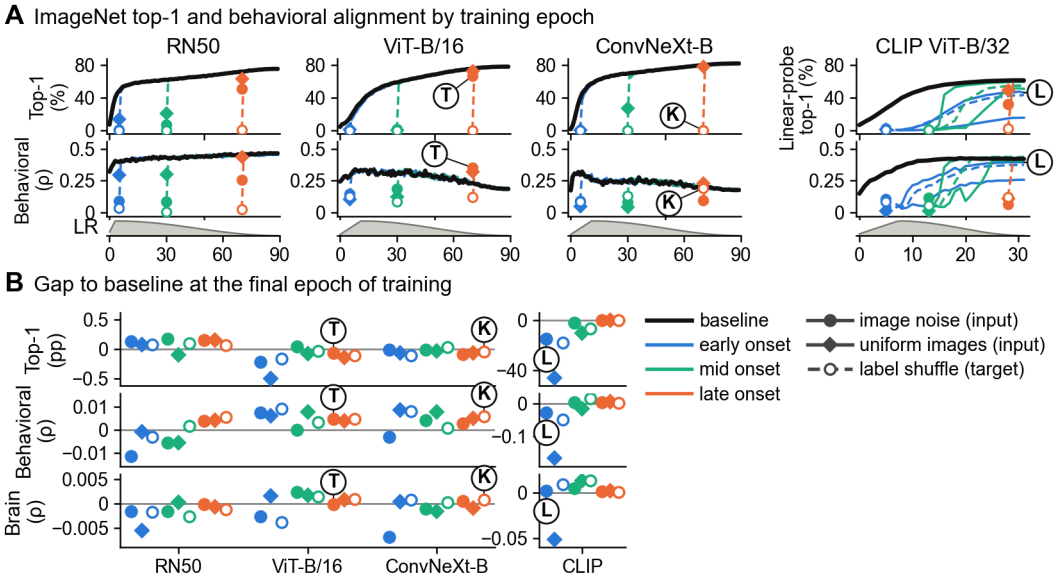

Figure 2: **Supervised classifiers recover from every perturbation; only CLIP ViT-B/32 keeps lasting deficits.** Color, onset (early, mid, late; epochs in Section 3.3); marker, perturbation type (filled circle, image noise; diamond, uniform images; open circle and dashed line, label shuffle). T, K, and L mark the runs tracked through the paper: ViT-B/16 with late image noise, ConvNeXt-B with late label shuffle, and CLIP ViT-B/32 with early image noise. Later figures use this encoding wherever they mark onsets, perturbation types, or tracked runs. RN50 uses re-estimated BatchNorm statistics throughout (Appendix E); pp, percentage points. (A) CLIP ViT-B/32 top-1 is from a linear probe; LR, baseline learning rate. (B) Final epoch: 89 (CLIP ViT-B/32: 31). Sub-epoch recovery and half-recovery times: Figure 6.

3.3 Perturbation experiment design

A perturbation has a *type* δ and an *onset* epoch t: every training batch in epoch t is modified by δ, and training is otherwise bit-identical to an unperturbed baseline run from the same initialization seed, so any lasting effect comes from the training trajectory, not from a persistently corrupted input or objective.

Two types corrupt the images: *image noise* replaces each input with standard Gaussian noise in the normalized input space, matching the first- and second-order statistics of normalized images while removing all natural structure, and *uniform images* sets every input to the constant 0.5. Two corrupt the supervision: *label shuffle* permutes the targets within each batch, so each image is paired with another image's target, and *random target* replaces each target with an independent Gaussian draw matching the per-dimension mean and standard deviation of the training targets. Random target requires continuous targets, so it is applied only to CLIP-HBA-Behavior.

Every run is perturbed for exactly one epoch. For fine-tuning, the onset sweeps a 28-point grid (every epoch from 0 to 9, then every fifth to 99) for all four types and three initialization seeds; following Zhao et al. (2025), fine-tuning allows up to 500 epochs but stops early with a patience of 20, and the baselines stop after 112–123 epochs, so the grid covers the first 100. From scratch, following Achille et al. (2019) and Lee et al. (2026), we sample three onsets per architecture (early, mid, late): epochs 5, 30, and 70 of 90 for the ImageNet classifiers, and 5, 13, and 28 of 32 for CLIP-ViT-B/32.

Baselines differ in shape: over the second half of training, alignment and ImageNet accuracy move in opposite directions in two of the four models, so the two can dissociate even without a perturbation (Appendix H). We compare each perturbed run with its baseline at the same epoch, as a difference or as a fraction of the baseline's value (e.g., Figure 3A).

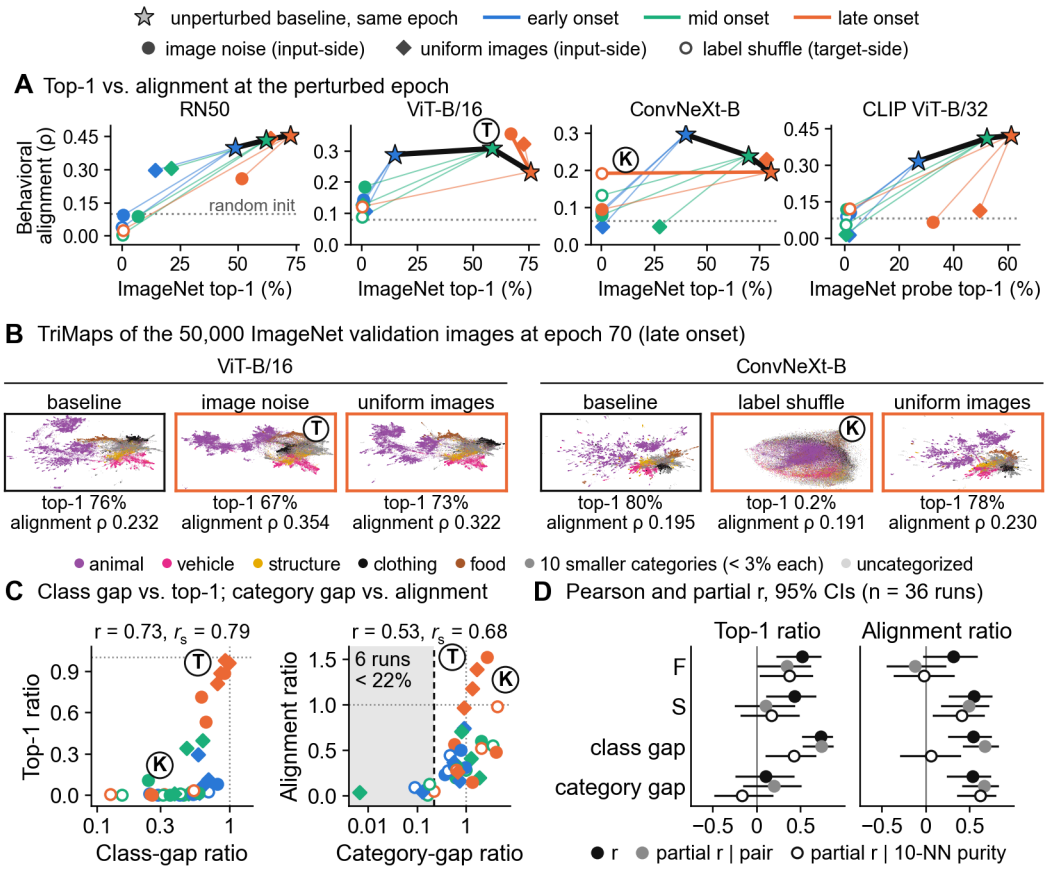

Figure 3: **Accuracy tracks class structure; alignment tracks category structure.** (A) Thin lines join each perturbed run to its baseline at the same epoch; black path, baseline across onsets; random init, mean of 5 untrained networks; CLIP ViT-B/32 top-1 from an ImageNet linear probe. (B) TriMaps of penultimate-layer features, colored by WordNet superordinate category and Procrustes-aligned to the baseline. (C) Class gap $(d_2 - d_1)/d_3$ and category gap $(d_3 - d_2)/d_3$, from mean-centered cosine distances between images of one class (d_1), of different classes in one category (d_2), and of different categories (d_3); ratios are perturbed over baseline; shaded, category gap below 22% of the baseline's. (D) Category fidelity F, Spearman ρ between perturbed and baseline RDMs of 16 category centroids; category separability S, cosine silhouette of the 16 categories. Partial r controls for the other metric of its pair or for 10-NN class purity; Fisher-z 95% CIs. Details in Appendix E.

4 Results

4.1 Perturbing one epoch has a range of immediate costs, shaped by onset timing, perturbation type, and training regime

One perturbed epoch lowers ImageNet top-1 by 2 to 80 percentage points (pp; Figure 2A). Early and mid, input-side perturbations cost the most: 13 of the 16 such runs keep at most 11% of their baseline's top-1, and the three exceptions are all uniform images (RN50 early and mid, 29% and 34%; ConvNeXt-B mid, 40%). Late, 7 of the 8 input-side runs keep 53–98%, while ConvNeXt-B with image noise falls to 0.2% (chance = 0.1%). Label shuffle drives top-1 to at most 1.9% at every onset in every architecture.

SPoSE alignment falls less, and early and mid onsets cost about the same (Figure 3A): averaged over architectures, input-side perturbations leave 0.34, 0.35, and 0.82 of the baseline's alignment at early, mid, and late onsets, and label shuffle 0.28, 0.25, and 0.46. Unlike for accuracy, which perturbation costs the most alignment depends on the architecture. Brain alignment shows the same timing (Figure 6C): input-side perturbations leave on average 0.48, 0.50, and 0.84 of it at early,

mid, and late onsets, and label shuffle 0.34, 0.39, and 0.55. It falls by up to 0.13 in the supervised classifiers and by 0.07 to 0.19 in CLIP ViT-B/32, and label shuffle costs it the most in 9 of the 12 architecture–onset pairs.

In ViT-B/16 and ConvNeXt-B, late input-side perturbations *raise* human alignment, except late image noise on ConvNeXt-B. ViT-B/16 with late image noise (case T) rises from $\rho = 0.232$ to 0.354, above its baseline's peak at any epoch (0.340), and ViT-B/16 and ConvNeXt-B with late uniform images rise to 0.322 and 0.230 ($+0.090$ and $+0.035$; Figure 3A), at a cost of 8.9, 3.1, and 2.0 pp of top-1. These three are the only runs of 36 whose brain alignment is at or above baseline at the perturbed epoch ($+0.026$, $+0.012$, and $+0.003$), and all three also exceed baseline under CKA (Appendix G). The gains appear where the baseline's alignment is falling: from 0.306 at epoch 30 to 0.232 at epoch 70 in ViT-B/16, and from 0.296 at epoch 5 to 0.195 in ConvNeXt-B, whereas in RN50 and CLIP ViT-B/32 it is still rising between the mid and late onsets (0.432 to 0.454; 0.410 to 0.423).

4.2 Task accuracy and human alignment can dissociate

Across the 36 runs, the fractions of top-1 and of behavioral alignment kept at the perturbed epoch correlate ($r = 0.67$), but only through runs that keep or lose both; the rank correlation is weak ($r_s = 0.32$, $p = 0.053$). The dissociation runs both ways (Figure 3A): K keeps 0.98 of its alignment with top-1 at 0.2%, and RN50 with early and mid uniform images keeps 0.75 and 0.71 of its alignment but 0.29 and 0.34 of its top-1, whereas CLIP ViT-B/32 with late uniform images keeps 0.81 of its top-1 but 0.26 of its alignment, and with late image noise 0.53 and 0.15. After each mid-training perturbation, CLIP ViT-B/32 ends training below baseline in accuracy but with its brain alignment recovered, and in two of the three cases its behavioral alignment too (Section 4.3).

The TriMaps in Figure 3B suggest what separates these outcomes. At epoch 70, both baselines arrange the images into many small islands, mostly single classes, grouped into regions by superordinate category (Appendix E). Late uniform images, which keep both accuracy and alignment, preserve this organization. In case T the animal islands merge into a few larger masses; in case K the islands dissolve into one diffuse cloud, yet images of a category still occupy the same part of it.

We quantify this organization in the features themselves. Category fidelity F asks whether the 16 category centroids keep their baseline arrangement, category separability S whether images lie closer to their own category than to the nearest other one, and 10-NN class purity whether an image's nearest neighbors share its class. Across the 36 runs, F predicts top-1 retention ($r = 0.52$) and S predicts alignment retention ($r = 0.55$; Figure 7); each relation survives controlling for the other measure (partial $r = 0.34$ and 0.49) and for 10-NN class purity (0.38 and 0.41), with 95% CIs excluding 0 (Figure 3D), while neither cross pair does ($|\text{partial } r| \leq 0.17$). Retained 10-NN class purity tracks top-1 retention almost exactly ($r = 0.95$) and alignment retention less closely ($r = 0.77$, $r_s = 0.51$). Accuracy thus depends on keeping the representation the classifier was trained on, whereas alignment depends more on categories remaining separable than on their arrangement.

To ask which distinctions a perturbed representation still draws, we compare mean-centered distances between images of the same class (d_1), of different classes in the same superordinate category (d_2), and of different categories (d_3). The class gap $(d_2 - d_1)/d_3$ measures how well classes are separated within a category, and the category gap $(d_3 - d_2)/d_3$ how well categories are separated beyond their classes. Relative to baseline, the category gap predicts alignment retention ($r_s = 0.68$) but not top-1 retention ($r_s = 0.13$), and the class gap predicts top-1 retention ($r_s = 0.79$) more than alignment retention ($r_s = 0.38$; Figure 3C, D). The six runs that keep the least alignment (0.01–0.13: RN50 with label shuffle at all three onsets, and CLIP ViT-B/32 with early and mid uniform images and mid label shuffle) are exactly the six whose category gap falls below 22% of baseline; their class gaps are also at most 0.073, so the three distances nearly converge. No other run's category gap falls below 37% of baseline.

Case K shows how a representation can lose accuracy but keep alignment. After one epoch of shuffled labels, ConvNeXt-B maps every image close to one shared vector (the mean feature vector is 19 times longer than the average residual, against 0.16 at baseline), so its classifier receives nearly the same input for every image. Behavioral alignment depends only on the rank order of distances and is blind to this uniform shrinkage; what remains is the residual, organized by category rather than class: K's class gap falls from 0.547 to 0.142, while its category gap grows 4.4-fold, from 0.024

to 0.104, the largest relative increase of any run. Sixteen runs collapse this way (mean within-class cosine similarity ≥ 0.95, against 0.29–0.58 at baseline). None keeps more than 25% of its 10-NN class purity, and the ten supervised ones keep at most 1.2% of their top-1, yet their alignment retention ranges from 0.13 to 0.98 and tracks S ($r = 0.57$, $p = 0.02$), not class purity ($r = -0.21$, $p = 0.43$).

The converse, accuracy kept but alignment lost, appears clearly only in CLIP ViT-B/32 at its late onset. With late uniform images it keeps 81% of its top-1, $F = 0.88$, and 44% of its 10-NN class purity, but only 0.26 of its alignment, as both gaps shrink ($\times 0.81$ and $\times 0.68$) and S falls from -0.017 to -0.130. Because its top-1 comes from a linear probe fit to the perturbed features, retained accuracy means only that class information is still linearly decodable, which can hold even when little of the geometry does (Figure 16C).

The category gap also explains why late perturbations raise alignment only in ViT-B/16 and ConvNeXt-B. In these two models, whose baseline alignment declines late in training, training trades category for class structure: between epochs 5 and 70 the class gap grows from 0.142 to 0.322 (ViT-B/16) and from 0.281 to 0.547 (ConvNeXt-B), while the category gap shrinks from 0.079 to 0.026 and 0.024. In RN50 and CLIP ViT-B/32, whose alignment keeps rising, the category gap stays between 0.079 and 0.093 at all three onsets. The three late perturbations that raised alignment partly reverse the trade without giving up class structure: the standard deviation of the class-centroid RDM, which training had lowered from 0.191 to 0.056 (ViT-B/16) and from 0.145 to 0.052 (ConvNeXt-B), rises in these runs by 79%, 27%, and 12%. Only two other runs raise it, ConvNeXt-B with mid and CLIP ViT-B/32 with early uniform images, and both lose their category structure and most of their alignment (0.20 and 0.04 of baseline).

Brain alignment agrees (Figure 14). The three runs that raised behavioral alignment are the only ones whose alignment with object-selective cortex rose: averaged over the nine LOC, FFA, EBA, and OFA ROIs, by $+0.049$ (T), $+0.026$, and $+0.015$, against at most $+0.004$ in every other run, most in right FFA and right EBA ($+0.086$ each in T). K's gains are confined to early visual cortex (V1 $+0.075$, V2 $+0.063$, V3 $+0.028$), while bilateral LOC falls by 0.024. Unrecognizable images that keep only mid-level texture and coarse form still evoke the animate–inanimate and big–small organization of occipitotemporal cortex (Long et al., 2018), so such features may separate superordinate categories without identifying the object, consistent with K keeping category- but not class-level structure.

CLIP-HBA-Behavior shows the same distinction in an interpretable space, since its outputs are the 66 SPoSE dimensions (Figure 5; Appendix C). At onsets from 14 on, random target and label shuffle shrink each dimension's variance across the 48 test images by 81% and 89% on average, for semantic dimensions such as "animal-related" and perceptual ones such as "metallic" alike, and cost 0.16 and 0.15 of alignment; input-side perturbations shrink it by 17–26% and cost 0.016–0.026. Early in fine-tuning (onsets 2–5), random target shrinks the variance by only 11–16% and raises alignment by up to 0.071. Ten epochs after the perturbed epoch, the variance under random target is back within 3% of the baseline's.

4.3 Supervised classifiers recover task performance, transfer, and human alignment

In the supervised classifiers, recovery is fast (Figure 6A, B). By the end of the next epoch, 26 of the 27 runs are within 3.3 pp of their baseline's top-1 at the onset (the baseline itself changes by -0.3 to $+5.6$ pp over that epoch), and each regains half its onset drop within 0.42 epoch; the exception, ConvNeXt-B with mid image noise, stays at chance ($\leq 0.12\%$) through that epoch and takes two epochs to regain half. Behavioral alignment recovers at least as fast. Recovery also undoes the late gains: T's behavioral alignment goes from $+0.121$ above baseline at the perturbed epoch to -0.001 one epoch later, although its representational distance to the baseline ($1-$ Spearman ρ between the two models' RDMs over 720 THINGS-fMRI concepts) is still 0.06 at the final epoch (Figure 8).

Linear-probe recovery mirrors the class–category split (Figure 4A). At the end of its perturbed epoch, K's 1000-class ImageNet probe is at chance (0.1%; baseline 80.2%), but its 16-category probe is at 66.3% (baseline 90.5%; majority class 39.8%). On transfer, K keeps 0.82 of its baseline's accuracy on three coarse 10-way datasets (CIFAR-10, STL-10, MNIST) but only 0.24 on the five fine-grained ones, whose classes are subordinate categories (chance-corrected). The asymmetry

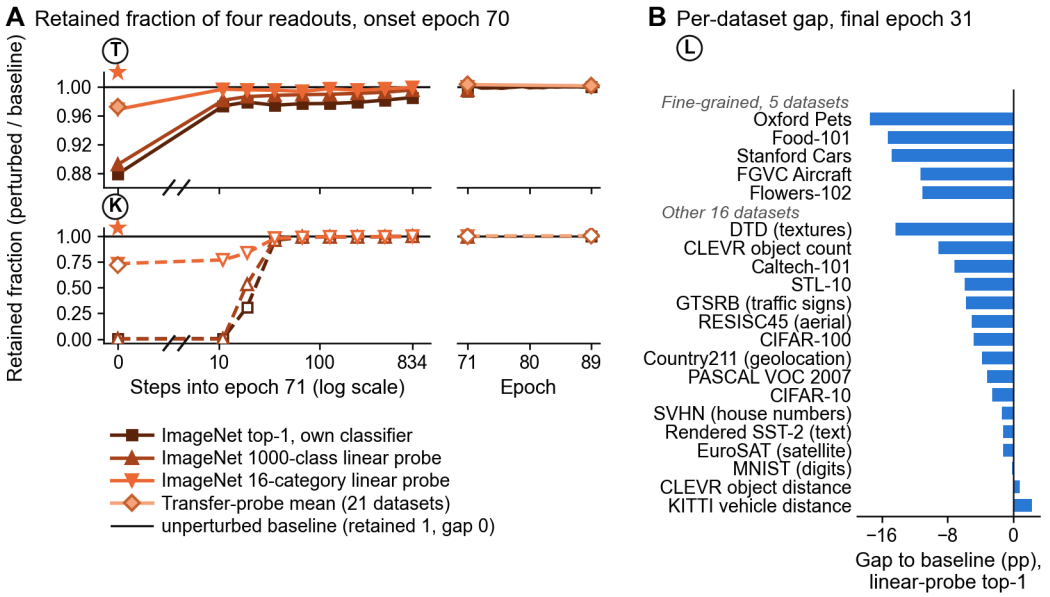

Figure 4: **Coarse transfer survives where fine-grained transfer fails.** Transfer-probe mean: mean linear-probe top-1 over the 21 datasets in B. (A) Star, end of perturbed epoch 70; the 16-category probe classifies WordNet superordinate categories. (B) Fine-grained datasets have subordinate-level classes. All 36 runs: Figure 9; protocol and scores: Appendix F.

is general: in all 32 runs in which either group falls below 0.9 of baseline at the onset, the fine-grained datasets lose more, and fine-grained retention tracks top-1 retention ($r = 0.85$) and 10-NN class purity ($r = 0.86$).

CLIP ViT-B/32 is the one model with lasting damage, and in it accuracy and alignment recover differently (Figure 2B). After all three mid-onset perturbations, brain alignment ends above baseline ($+0.004$ to $+0.013$), and behavioral alignment ends above it after image noise and label shuffle ($+0.002$ and $+0.015$) but below after uniform images (-0.016). Alignment also regains half its onset drop sooner: in 5, 11, and 8 epochs after the early perturbations and 3, 10, and 5 after the mid ones, against 11, 23, and 11 and 3, 11, and 7 for top-1 (Figure 6B). Early uniform images is the one perturbation from which CLIP ViT-B/32 recovers no readout; its representation stays collapsed (Section 5.1; Figures 8 and 16).

CLIP ViT-B/32's lasting losses fall on fine-grained information (Figure 4B). In each of its six runs with a lasting top-1 deficit, the fine-grained datasets keep less of their baseline's accuracy at the final epoch than the coarse ones (0.31–0.97 against 0.79–1.00); for L, the five fine-grained datasets end 14.0 pp below baseline on average, against 3.9 pp for the other 16. The mid-onset runs, whose alignment recovered, keep nearly all their coarse category information (0.98–1.00) while still missing fine-grained information (0.84–0.97), consistent with alignment reading category-level and accuracy class-level structure. Transfer degrades less than ImageNet accuracy (after early uniform images, CLIP ViT-B/32 keeps 0.66 of its transfer mean but 0.25 of its probe top-1; Figure 9A), and across all 108 checkpoints (36 runs at the onset, onset $+1$, and final epochs) the transfer gap grows with representational distance from the baseline ($r_s = -0.70$; -0.62 without CLIP ViT-B/32; Figure 9B).

5 Discussion

5.1 Interpretation

On average, early and mid perturbations cause the largest immediate disruption. CLIP ViT-B/32 is the only model with lasting damage, and there it depends on onset as a critical period would: all six of its runs with a lasting accuracy deficit have early or mid onsets, none a late one. The clearest

case is early uniform images, which collapse the representation onto a single axis that is still one-dimensional five epochs later and has regained only 14.2 effective dimensions by the end of training, against 115.5 for the baseline; the same perturbation late in training leaves no lasting deficit. We take this critical period to be real but, so far, shown in a single model (Section 5.2). ImageNet accuracy depends mainly on class-level structure and human alignment mainly on category-level structure, so a perturbation that destroys only the former costs accuracy without costing alignment, and one that restores the latter can raise alignment at little cost to accuracy. This parallels Muttenthaler et al. (2023), who found that aligning global structure with human judgments harms downstream performance unless local structure is preserved. Recovering the task brought alignment back in all 27 supervised runs and all three late CLIP ViT-B/32 runs, but not conversely: after mid-training perturbations, CLIP ViT-B/32 recovers its alignment without its accuracy.

Human perception is highly robust. Humans can classify objects across a wide range of corruptions and transformations, often quickly and with little effort (Geirhos et al., 2018), and aligning models with human vision may make them more robust without massive amounts of training time and data (Sucholutsky & Griffiths, 2023; Muttenthaler et al., 2025). If the objective is human alignment, the standard critical-period intuition transfers only partially: it largely holds for ImageNet accuracy, where data quality matters most early in training, but timing that suits the task objective may be premature for another. Critical periods should therefore be identified for each objective of interest, not for task accuracy alone.

5.2 Limitations and Future Directions

Due to computational limits, the four from-scratch models were trained with one seed each, and more seeds would give a more reliable picture of the effects. The lasting effects also rest on a narrow base. Three of the four from-scratch models recover fully from every perturbation, usually within a fraction of an epoch. Additionally, the lasting deficits appear only one model, CLIP ViT-B/32. Its perturbations cover a larger proprtion of its total training (one epoch is 3.1% of its 32-epoch schedule, against 1.1% of the 90-epoch ImageNet schedules), and its early onset (epoch 5) falls inside its learning-rate warmup (2,500 steps, about 8 epochs). Warmup alone does not explain the deficits. The early onsets of ViT-B/16 and ConvNeXt-B also fall inside their warmup (10,000 steps, about 12 epochs) and both recover fully, while CLIP ViT-B/32's mid onset (epoch 13) comes after warmup and still leaves a lasting accuracy deficit. But CLIP ViT-B/32 also differs from the classifiers in objective, data, and schedule, and our design cannot say which difference produces its deficits.

Finally, the training recipes in our five models are not matched. For example, RN50 used stochastic gradient descent while the others used AdamW. Researchers may alternatively choose to examine the effects of perturbations under matched training paradigms. Our hope is that these initial results provide a framework under which future research can examine more closely how particular features such as network depth, information plasticity, data recipes, or optimization techniques influence the effect of perturbations and presence of a critical period with regard to human representational alignment.

REFERENCES

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018. URL http://jmlr.org/papers/v19/17-646.html.

Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical learning periods in deep networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BkeStsCcKQ.

Michael S. Brainard and Eric I. Knudsen. Sensitive periods for visual calibration of the auditory space map in the barn owl optic tectum. *Journal of Neuroscience*, 18(10):3929–3942, 1998. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.18-10-03929.1998.

Joel Dapello, Kohitij Kar, Martin Schrimpf, Robert Geary, Michael Ferguson, David D. Cox, and James J. DiCarlo. Aligning model and macaque inferior temporal cortex representations improves model-to-human behavioral alignment and adversarial robustness. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=SMYdcXjJh1q.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16×16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2010.11929.

Thomas Fel, Ivan F. Rodriguez Rodriguez, Drew Linsley, and Thomas Serre. Harmonizing the object recognition strategies of deep neural networks with humans. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9432–9446, 2022. doi: 10.52202/068431-0685.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems*, volume 36, pp. 50742–50768, 2023. doi: 10.52202/075280-2208.

Vinicius Yuiti Fukase, Heitor Gama, Barbara Bueno, Lucas Libanio, Anna Helena Reali Costa, and Artur Jordao. One period to rule them all: Identifying critical learning periods in deep networks. *Procedia Computer Science*, 264:270–279, 2025. doi: 10.1016/j.procs.2025.07.138.

Robert Geirhos, Carlos R. M. Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge, and Felix A. Wichmann. Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/0937fb5864ed06ffb59ae5f9b5ed67a9-Abstract.html.

Abdulkadir Gokce and Martin Schrimpf. Scaling laws for task-optimized models of the primate visual ventral stream. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 19712–19733. PMLR, 2025. URL https://proceedings.mlr.press/v267/gokce25a.html.

Zhengxin Gong, Ming Zhou, Yuxuan Dai, Yushan Wen, Youyi Liu, and Zonglei Zhen. A large-scale fMRI dataset for the visual processing of naturalistic scenes. *Scientific Data*, 10(1):559, August 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-02471-x.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Martin N. Hebart, Adam H. Dickter, Alexis Kidder, Wan Y. Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I. Baker. THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, 14(10):e0223792, 2019. doi: 10.1371/journal.pone.0223792.

Takao K. Hensch. Critical period regulation. *Annual Review of Neuroscience*, 27:549–579, 2004. doi: 10.1146/annurev.neuro.27.070203.144327.

David H. Hubel and Torsten N. Wiesel. The period of susceptibility to the physiological effects of unilateral eye closure in kittens. *Journal of Physiology*, 206(2):419–436, 1970. doi: 10.1113/jphysiol.1970.sp009022.

Philipp Kaniuth and Martin N. Hebart. Feature-reweighted representational similarity analysis: A method for improving the fit between computational models, brains, and behavior. *NeuroImage*, 257:119294, 2022. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2022.119294.

Bidur Khanal, S. M. Kamrul Hasan, Bishesh Khanal, and Cristian A. Linte. Investigating the impact of class-dependent label noise in medical image classification. *Proceedings of SPIE–the International Society for Optical Engineering*, 12464:1246437, February 2023. ISSN 0277-786X. doi: 10.1117/12.2654420.

Michael Kleinman, Alessandro Achille, and Stefano Soatto. Critical learning periods emerge even in deep linear networks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Aq35gl2c1k.

Eric I. Knudsen. Sensitive periods in the development of the brain and behavior. *Journal of Cognitive Neuroscience*, 16(8):1412–1425, October 2004. doi: 10.1162/0898929042304796.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019a.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better ImageNet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2661–2671, 2019b.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

Manoj Kumar, Neil Houlsby, Nal Kalchbrenner, and Ekin Dogus Cubuk. Do better ImageNet classifiers assess perceptual similarity better? *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=qrGKGZZvH0.

JangHyeon Lee, Philipe Ambrozio Dias, Yao-Yi Chiang, and Dalton Lunga. Understanding the learning phases in self-supervised learning via critical periods. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=UxIRc97ecL.

Drew Linsley, Ivan F. Rodriguez Rodriguez, Thomas Fel, Michael Arcaro, Saloni Sharma, Margaret Livingstone, and Thomas Serre. Performance-optimized deep neural networks are evolving into worse models of inferotemporal visual cortex. In *Advances in Neural Information Processing Systems*, volume 36, pp. 28873–28891, 2023. doi: 10.52202/075280-1257.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32100–32121. PMLR, 2024. URL https://proceedings.mlr.press/v235/liu24bn.html.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022. doi: 10.1109/CVPR52688.2022.01167.

Bria Long, Chen-Ping Yu, and Talia Konkle. Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences*, 115 (38):E9015–E9024, 2018. doi: 10.1073/pnas.1719616115.

Peter Marler. A comparative approach to vocal learning: Song development in white-crowned sparrows. *Journal of Comparative and Physiological Psychology*, 71(2, Pt. 2):1–25, May 1970. doi: 10.1037/h0029144.

Lukas Muttenthaler, Lorenz Linhardt, Jonas Dippel, Robert A. Vandermeulen, Katherine Hermann, Andrew K. Lampinen, and Simon Kornblith. Improving neural network representations using human similarity judgments. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 50978–51007, 2023. doi: 10.52202/075280-2218.

Lukas Muttenthaler, Klaus Greff, Frieda Born, Bernhard Spitzer, Simon Kornblith, Michael C. Mozer, Klaus-Robert Müller, Thomas Unterthiner, and Andrew K. Lampinen. Aligning machine and human visual representations across abstraction levels. *Nature*, 647(8089):349–355, November 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09631-6.

Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 23296–23308, 2021. URL https://arxiv.org/abs/2105.10497.

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9616–9625, 2019. doi: 10.1109/ICCV.2019.00971.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://arxiv.org/abs/2103.00020.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Ilia Sucholutsky and Thomas L. Griffiths. Alignment with human representations supports robust few-shot learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 73464–73479, 2023. doi: 10.52202/075280-3213.

Shobhita Sundaram, Stephanie Fu, Lukas Muttenthaler, Netanel Y. Tamir, Lucy Chai, Simon Kornblith, Trevor Darrell, and Phillip Isola. When does perceptual alignment benefit vision representations? In *Advances in Neural Information Processing Systems*, volume 37, pp. 55314–55341, 2024. doi: 10.52202/079017-1756.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. doi: 10.1145/2812802.

Janet F. Werker and Richard C. Tees. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7(1):49–63, 1984. ISSN 0163-6383. doi: 10.1016/S0163-6383(84)80022-3.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. A large-scale study of representation learning with the visual task adaptation benchmark, 2020. URL https://arxiv.org/abs/1910.04867.

Stephen Chong Zhao, Yang Hu, Jason Lee, Andrew Bender, Trisha Mazumdar, Mark Wallace, and David A. Tovar. Shifting attention to you: Personalized brain-inspired AI models, 2025. URL https://arxiv.org/abs/2502.04658.

6 AI USE STATEMENT

In this work, we used generative AI tools to implement methods, to provide feedback on research methodology and experiments, and to help clean datasets. We have not used generative AI tools to generate synthetic data sets. Developing theoretical models or conceptual frameworks, formulating mathematical claims, proving mathematical claims, translation, and qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to create and edit software code, to create and modify scientific figures, to create artifacts (the released code repository), and to edit the paper for readability. We have reviewed all AI-assisted work. AI-written code was reviewed and tested by the authors, the released code was verified to reproduce every figure and table in the paper, and every number and claim in the text was checked against the underlying results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All code, configuration files, and aggregated results needed to reproduce our findings are available at an anonymous repository (https://anonymous.4open.science/r/temporal-perturbations). The repository contains the training code for all four from-scratch model families and for CLIP-HBA-Behavior fine-tuning, the exact training arguments of every baseline and perturbed run, and the evaluation code for every alignment, accuracy, and transfer measure. It also contains the script that generates each figure and table and the reference data we derived from public datasets: the SPoSE embeddings and the THINGS-fMRI and NOD region-of-interest RDMs. Because the aggregated evaluation results are included, every figure, table, and number in the Results can be regenerated on a CPU without the datasets or checkpoints; the high-dimensional features behind the embedding-map panels are not included, so those panels are re-drawn from saved two-dimensional coordinates. Model architectures, datasets, and the perturbation design are described in Section 3, and training hyperparameters for every model family are listed in Appendix A (Tables 1–5). Descriptions of the stimulus sets and CLIP-HBA-Behavior are in Appendix B, its fine-tuning results in Appendix C, and the alignment metrics are defined in Section 3.2 and Appendix D. The evaluation protocol behind each main-text figure is in Appendix E, and the linear-probe protocol, per-dataset scores, and uncertainty estimates are in Appendix F. Robustness to the choice of alignment metric and replication of the neural results on a second fMRI dataset are reported in Appendices G and I.

A TRAINING CONFIGURATIONS

Table 1: Hyperparameters used to train CLIP-HBA-Behavior. The CLIP ViT-L/14 backbone is frozen; only the DoRA adapters are trained against 66-dimensional SPoSE behavioral targets.

Hyperparameter	Value
Architecture	
Image encoder	ViT-L/14 (CLIP, OpenAI; frozen)
Text encoder	Transformer (CLIP; frozen)
Patch size	14
Input resolution	224×224
Adaptation	DoRA (weight-decomposed low-rank)
DoRA rank r	32
DoRA α	16
DoRA dropout	0.1
Adapted layers	last 2 vision blocks + last 1 text block
Readout / output	66-d SPoSE embedding (image–prompt similarity)
Parameters (total)	427.8M
Trainable parameters	0.19M (DoRA only; 0.04%)
Optimization	
Optimizer	AdamW
Adam β_1, β_2	0.9, 0.999
Adam ϵ	1×10^{-8}
Learning rate	3×10^{-4}
LR schedule	constant
Warmup	none
Weight decay	0
Loss	mean squared error (66-d SPoSE targets)
Batch size (global)	64
Gradient accumulation	1
Epochs	500 (max)
Early stopping	patience 20 (validation MSE)
Gradient clipping	None
Data & Augmentation	
Dataset	THINGS object images
Targets	66-d SPoSE behavioral embeddings (rescaled)
Training set	1,806 images, 80/20 train/val split ($\sim$1,444/362)
Behavioral-RSA eval set	48 held-out images (fixed order)
Normalization	THINGS mean $= [0.530, 0.481, 0.419]$, std $= [0.276, 0.266, 0.282]$
Augmentation	none (resize to 224×224 only)
System	
Hardware	$1\times$ NVIDIA A100 (40 GB) or RTX A6000 (48 GB)
Distributed strategy	none (single GPU)
Mixed precision	none (FP32)
Random seed(s)	3 (seeds 1–3; each perturbed run forks from the same-seed baseline)

Table 2: Pretraining hyperparameters for CLIP ViT-B/32 on YFCC15M.

Hyperparameter	Value
Architecture	
Image encoder	ViT-B/32
Text encoder	Transformer
Patch size	32
Input resolution	224×224
Embedding (projection) dim.	512
Vision depth / width / heads	12 / 768 / 12
Text depth / width / heads	12 / 512 / 8
MLP ratio	4.0
Context length	77
Vocabulary size	49,408
Activation	GELU
Learnable temperature (init)	$\ln(100) \approx 4.605$, clamped ≤ 100
Parameters (total)	151.3M
Trainable parameters	151.3M (all)
Optimization	
Optimizer	AdamW
Adam β_1, β_2	0.9, 0.98
Adam ϵ	1×10^{-8}
Learning rate	5×10^{-4} (peak)
LR schedule	cosine decay (min 0)
Warmup	2,500 steps
Weight decay	0.1
Loss	InfoNCE (contrastive)
Batch size (global)	49,152
Batch size (per GPU)	1,024
Gradient accumulation	6
Epochs	32
Early stopping	none
Gradient clipping	None
Data & Augmentation	
Dataset	YFCC15M ($\sim$15M image–text pairs)
Normalization	CLIP mean/std
RandomResizedCrop scale	$(0.2, 1.0)$
ColorJitter (b, c, s, h), prob.	$(0.4, 0.4, 0.4, 0.1)$, p=0.8
RandomGrayscale prob.	0.2
GaussianBlur σ, prob.	$[0.1, 2.0]$, p=0.5
RandomHorizontalFlip prob.	0.5
System	
Hardware	$8\times$ A100-SXM4 (80 GB)
Distributed strategy	DDP
Mixed precision	bf16
Random seed(s)	1 (seed 42)

Table 3: Training hyperparameters for ResNet50 on ImageNet-1k.

Hyperparameter	Value
Architecture	
Backbone	ResNet50 (timm `resnet50`)
Input resolution	224×224
Classifier head	linear, 1000-way (ImageNet-1k)
Parameters (total)	25.6M
Trainable parameters	25.6M (all; trained from scratch)
Optimization	
Optimizer	SGD (Nesterov momentum)
Momentum	0.9
Learning rate	0.4 (peak)
LR schedule	cosine decay (min 0)
Warmup	5,000 steps
Weight decay	1×10^{-4}
Loss	cross-entropy, label smoothing 0.1
Batch size (global)	1,024
Batch size (per GPU)	256
Gradient accumulation	1
Epochs	90
Gradient clipping	None
Data & Augmentation	
Dataset	ImageNet-1k (1,281,167 train / 50,000 val)
Normalization	standard ImageNet mean/std
Interpolation	bilinear
Mixup / CutMix	none
RandAugment	none
System	
Hardware	$4\times$ A100 (40 GB)
Distributed strategy	DDP, sync BatchNorm (single node)
Mixed precision	none (FP32)
Random seed(s)	42

Table 4: Training hyperparameters for ConvNeXt-B on ImageNet-1k.

Hyperparameter	Value
Architecture	
Backbone	ConvNeXt-B (timm `convnext_base`)
Input resolution	224×224
Classifier head	linear, 1000-way (ImageNet-1k)
Parameters (total)	88.6M
Trainable parameters	88.6M (all; trained from scratch)
Optimization	
Optimizer	AdamW
Adam β_1, β_2	0.9, 0.999
Adam ϵ	1×10^{-8}
Learning rate	0.004 (peak)
LR schedule	cosine decay (min 1×10^{-6})
Warmup	10,000 steps
Weight decay	0.05
Loss	cross-entropy, label smoothing 0.1
Batch size (global)	1,536
Batch size (per GPU)	256
Gradient accumulation	1
Epochs	90
Gradient clipping	1.0 (norm)
Data & Augmentation	
Dataset	ImageNet-1k (1,281,167 train / 50,000 val)
Normalization	standard ImageNet mean/std
Interpolation	bicubic
RandAugment	rand-m9-mstd0.5-inc1
Mixup α, prob., switch prob.	0.8, p=1.0, p=0.5 (batch mode)
CutMix α	1.0
System	
Hardware	$6\times$ A100 (40 GB)
Distributed strategy	DDP, sync BatchNorm (single node)
Mixed precision	bf16
Random seed(s)	42

Table 5: Training hyperparameters for ViT-B/16 on ImageNet-1k.

Hyperparameter	Value
Architecture	
Backbone	ViT-B/16 (timm `vit_base_patch16_224`)
Patch size	16
Input resolution	224×224
Classifier head	linear, 1000-way (ImageNet-1k)
Parameters (total)	86.6M
Trainable parameters	86.6M (all; trained from scratch)
Optimization	
Optimizer	AdamW
Adam β_1, β_2	0.9, 0.999
Adam ϵ	1×10^{-8}
Learning rate	0.003 (peak)
LR schedule	cosine decay (min 1×10^{-6})
Warmup	10,000 steps
Weight decay	0.05
Loss	cross-entropy, label smoothing 0.1
Batch size (global)	1,536
Batch size (per GPU)	256
Gradient accumulation	1
Epochs	90
Gradient clipping	1.0 (norm)
Data & Augmentation	
Dataset	ImageNet-1k (1,281,167 train / 50,000 val)
Normalization	standard ImageNet mean/std
Interpolation	bicubic
RandAugment	rand-m9-mstd0.5-inc1
Mixup α, prob., switch prob.	0.8, p=1.0, p=0.5 (batch mode)
CutMix α	1.0
System	
Hardware	$6\times$ A100 (40 GB)
Distributed strategy	DDP (single node)
Mixed precision	bf16
Random seed(s)	42

B Model and Dataset Descriptions

B.1 THINGS

SPoSE embeddings are created by taking discrete choice judgments from a triplet odd-one-out task and optimizing a non-negative feature matrix to predict these choices. Each of the 66 resulting dimensions corresponds to a semantically meaningful perceptual feature (e.g., "metallic," "animal-related"), and together they reflect the primary features that humans use to understand their visual world. The 1,854 images with SPoSE embeddings therefore serve as our evaluation set for model–human representational alignment.

B.2 CLIP-HBA-Behavior

In CLIP-HBA-Behavior, the SPoSE text dimensions (e.g., "metallic," "animal-like") are fed into CLIP's text encoder, while the THINGS images are fed into its vision encoder. The dot-product projection of these two modalities produces a predicted 66-dimensional embedding per image, which is optimized against the ground-truth SPoSE embedding via mean squared error (MSE) loss using weight-decomposed low-rank adaptation (DoRA) (Liu et al., 2024). As described, the THINGS dataset from Hebart et al. (2019) has 1,854 images for which SPoSE embeddings are available.

Of these, 48 images were fully-sampled in the triplet odd-one-out task, while the remaining 1,806 images were only partially sampled due to practical constraints during experimentation. Therefore, CLIP-HBA-Behavior is fine-tuned on the 1,806 non-fully-sampled THINGS images using an 80/20 train/validation split; and the model is evaluated on the 48 fully-sampled images, which are held out during fine-tuning.

C CLIP-HBA-BEHAVIOR FINE-TUNING RESULTS

Behavioral alignment of CLIP-HBA-Behavior (Figure 5B) is the Spearman ρ between the model RDM ($1-$ Pearson r between 66-d outputs for 48 THINGS images) and the human RDM from triplet judgments on the same 48 objects. The variance change in Figure 5C is $\Delta\mathrm{Var} = (\mathrm{Var}(\mathrm{perturbed}) - \mathrm{Var}(\mathrm{baseline}))/\mathrm{Var}(\mathrm{baseline})$, where Var is a dimension's variance across the 48 images and the baseline shares the seed; rows are sorted by baseline variance within each group. Onsets 0–1 have one seed, except uniform images.

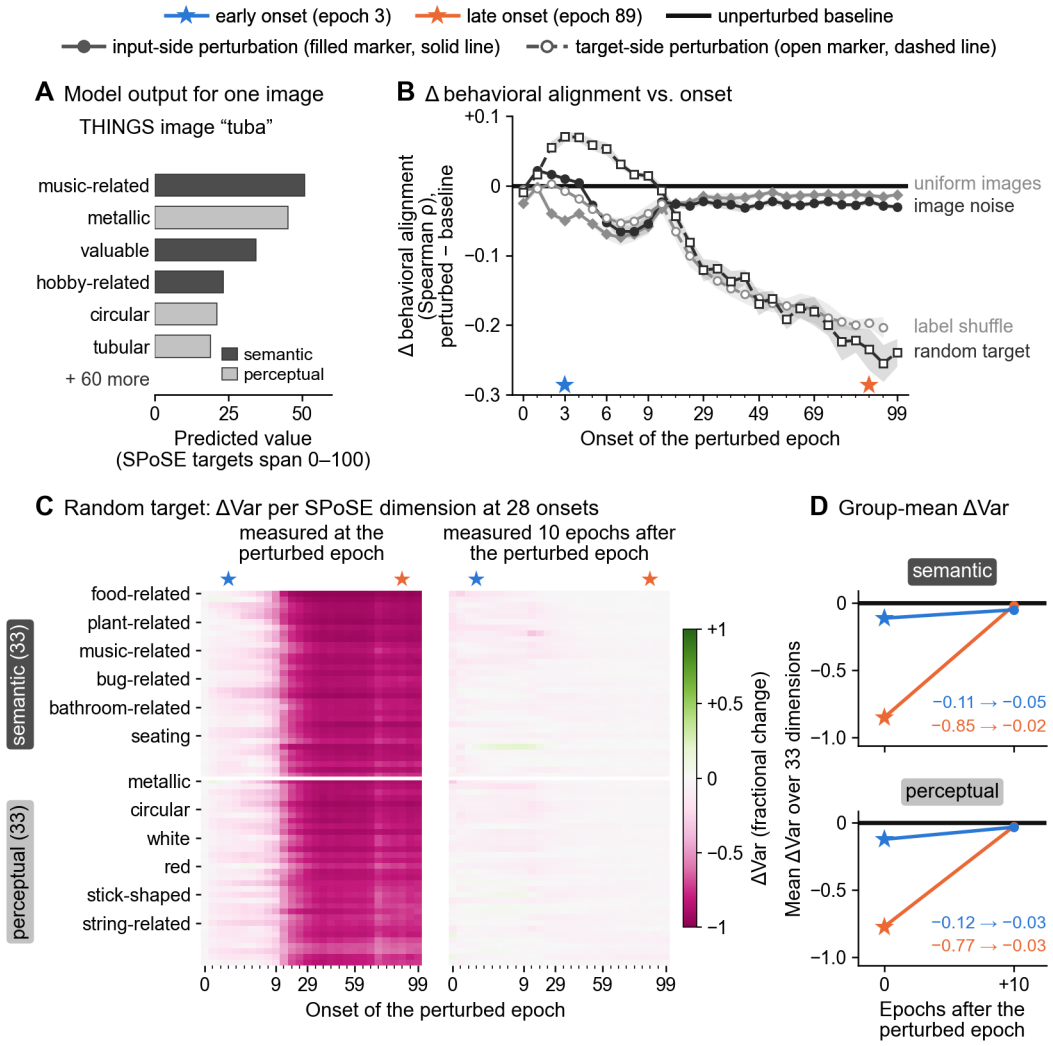

Figure 5: **Random target flattens perceptual and semantic dimensions alike, and both recover within 10 epochs.** CLIP-HBA-Behavior (CLIP ViT-L/14 with DoRA adapters) is fine-tuned to predict the 66 SPoSE dimensions; the 33 describing color, shape, texture, or material are grouped as perceptual, the rest as semantic. (A) Unperturbed model, seed 1, epoch 99. (B) Alignment at the end of the perturbed epoch against the same-seed baseline, on 48 THINGS images; bands, ±s.e.m. over 3 seeds. Onsets in B and C are every epoch from 0 to 9, then every fifth to 99, so the axes are not evenly spaced. (C) Relative change in each dimension's variance across the 48 images; mean of 3 seeds. (D) Group means of C; error bars, ±s.e.m. over the 33 dimensions.

D Alignment Metric Details

RSA and CKA. Rather than comparing individual embeddings directly, RSA operates on RDMs, which capture the full pairwise geometry of a representational space. CKA instead compares two representations directly in feature space, rather than through the rank structure of their pairwise dissimilarities. The two measures differ in what they are invariant to: RSA with Spearman correlation is invariant to any monotonic transformation of the RDM entries, and is therefore sensitive only to the rank order of pairwise dissimilarities, whereas CKA is invariant to orthogonal transformations and isotropic scaling of the raw embeddings. Reporting both reduces the likelihood that alignment effects reflect artifacts of a single similarity measure.

Layer-reweighted RSA. Rather than using penultimate-layer embeddings, neural comparison considers representations from many layers of the model. LR-RSA builds an RDM from every residual block, with convolutional blocks flattened over $C \times H \times W$ to preserve retinotopic layout and transformer blocks reduced to the CLS token. It then fits a non-negative ridge regression that reweights the per-layer RDMs to predict each ROI's RDM, with the ridge penalty selected by leave-conditions-out cross-validation. The 14 NOD ROIs are V1–V4, V8, LO, PIT, FFC, VMV, and VVC.

E Evaluation Details for the Main-Text Figures

BatchNorm re-estimation (RN50). A perturbed epoch corrupts RN50's BatchNorm running statistics, so its logged top-1 falls to $\leq 0.4\%$ at each onset whatever its weights have learned. Every RN50 value (top-1, behavioral and brain alignment, linear probes, and representational distances) is therefore computed after re-estimating the BatchNorm statistics on 8,000 clean ImageNet training images, for both the perturbed run and its baseline.

Accuracy. Top-1 is scored with each model's own classifier on the 50,000 ImageNet validation images; CLIP ViT-B/32 has no classifier and is scored with a logistic probe on its 768-d class-token features. Checkpoints within the epoch after the perturbed one (Figure 6A, B; Figure 4A, step axis) are scored on a fixed 10,000-image validation subset (10 images per class). That epoch is 1,251 optimizer steps for RN50 and 834 for the other models, and the first checkpoint follows 11 steps. On the step axis of Figure 4A, step 0 is compared with the baseline at epoch 70 and later checkpoints with the baseline at the end of epoch 71.

Half-recovery time (Figure 6B). The onset drop is the baseline minus the perturbed value at the onset epoch. Supervised values are interpolated between checkpoints, so values below about 0.01 epoch mean only that half the drop was regained before the first checkpoint; CLIP ViT-B/32 values are the first evaluated epoch, an upper bound. An arrow marks a run that had not regained half by the last checkpoint (1 epoch), a lower bound. Runs whose alignment fell by less than 0.01, about the baselines' median change between consecutive epochs, or rose, are marked "none".

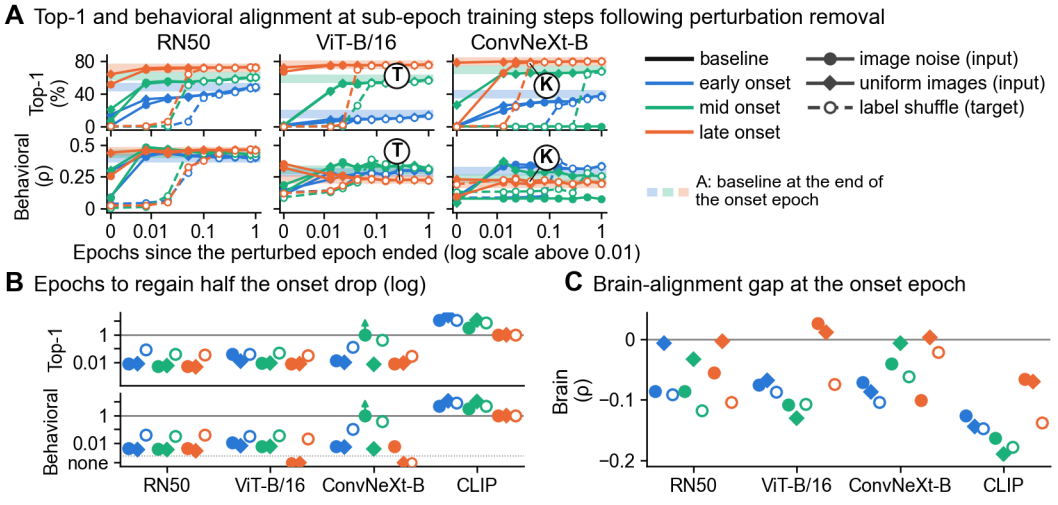

Figure 6: **Supervised classifiers usually regain half their loss within an epoch; CLIP ViT-B/32 takes up to 23 epochs.** Alignment as in Figure 2; RN50 uses re-estimated BatchNorm statistics throughout. (A) Checkpoints within the epoch after the perturbed one; CLIP ViT-B/32 has none. (B) Arrow, not regained within 1 epoch; none, alignment fell by less than 0.01 or rose. (C) Perturbed minus baseline at the end of the perturbed epoch.

Category structure (Figures 3 and 7). The 16 categories are the 15 WordNet superordinate categories plus uncategorized. Fidelity F and separability S are computed from penultimate-layer

features (CLIP ViT-B/32: 768-d class-token features) of the 50,000 validation images at each run's perturbed epoch; S ranges from -1 to 1, higher meaning more separated. Procrustes alignment allows rotation, reflection, and uniform scaling.

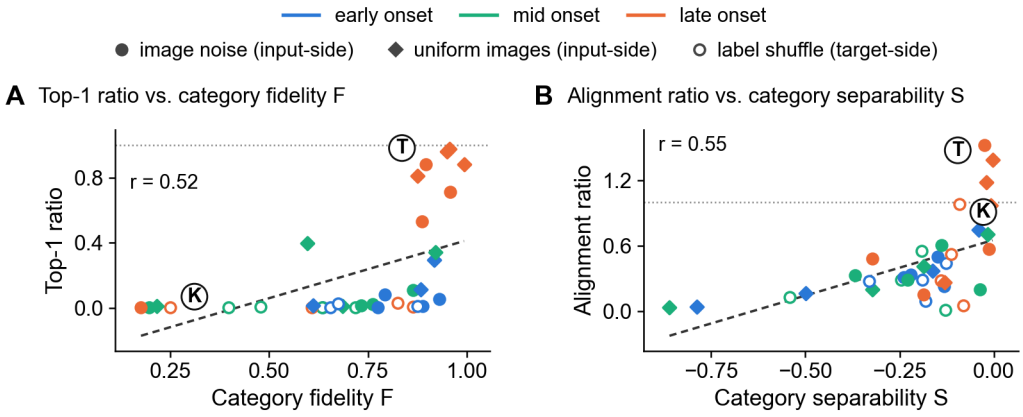

Figure 7: **Category fidelity predicts how much accuracy a run keeps; category separability predicts how much alignment it keeps.** Top-1 and alignment ratios, F, and S as in Figure 3; one point per perturbed run. r, Pearson; dashed lines, least-squares fits.

Distances at three levels of the category hierarchy (Figure 3C). For each run and its baseline at the perturbed epoch, we subtract the mean penultimate-layer feature vector over the 50,000 validation images and average the cosine distance over 20,000 random image pairs at each of three levels: the same ImageNet class (d_1), different classes in the same WordNet category (d_2), and different categories (d_3). Pairs involving uncategorized classes are excluded from the last two levels, and a run and its baseline share the same pairs. The class gap is $(d_2 - d_1)/d_3$ and the category gap $(d_3 - d_2)/d_3$; correlations across runs use each gap as a fraction of its baseline's. Without the mean subtraction, a representation that maps every image close to one shared vector has raw distances near 0 at every level even though the geometry of the residual around that vector survives. The object-selective ROIs of Section 4.2 are bilateral LOC and left and right LOC, FFA, EBA, and OFA.

TriMaps and representational distance. Each TriMap frame is a separate fit to penultimate-layer features; in Figure 8A, $\leq 3\%$ of points fall outside the shared axes. Representational distance compares RDMs built from penultimate-layer features (CLIP ViT-B/32: 768-d class-token features) for the 720 THINGS-fMRI concepts.

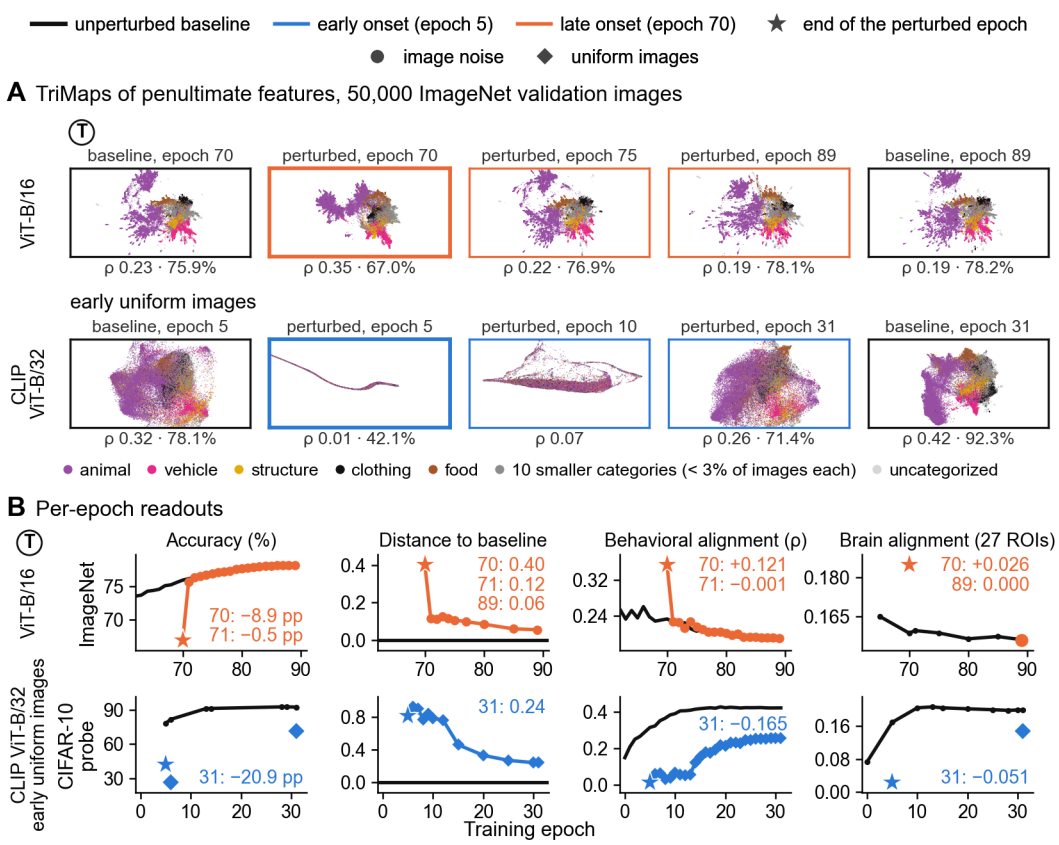

Figure 8: **Case T is back at baseline one epoch later; CLIP ViT-B/32 recovers only partly from early uniform images.** (A) TriMap embeddings as in Figure 3B, each frame aligned to the first in its row; below each, behavioral ρ and accuracy (not probed for the CLIP ViT-B/32 epoch-10 frame). (B) Accuracy is ImageNet top-1 for ViT-B/16 and CIFAR-10 linear-probe top-1 for CLIP ViT-B/32 (probed at the marked epochs). Distance to baseline: $1-$ Spearman ρ between the perturbed and baseline RDMs over 720 THINGS-fMRI concepts. Annotations, perturbed $-$ baseline at the listed epochs (distance: its value).

Transfer probes (Figures 4 and 9). Datasets, probe protocol, all scores, and their uncertainty are in Appendix F.

F Linear-Probe Transfer: Protocol, Scores, and Uncertainty

This section documents the probes behind Figures 4 and 9. All of them follow one recipe. Every probe record (dataset, checkpoint, top-1 and top-5 accuracy, selected C, selection settings, iterations, and convergence) is in the supplementary file `transfer_probe_scores.csv`.

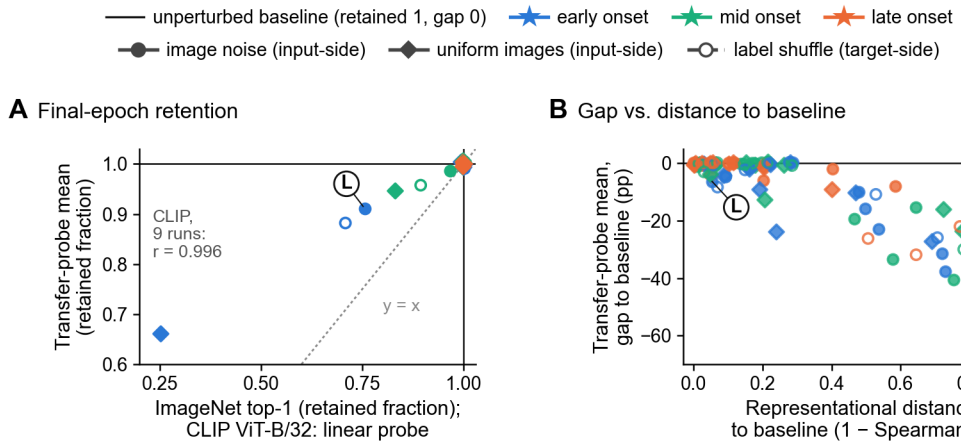

Figure 9: **Transfer degrades less than ImageNet accuracy, and its gap grows with distance from the baseline.** Transfer-probe mean as in Figure 4. (A) Final epoch: 89; CLIP ViT-B/32: 31. ImageNet top-1 is scored with each model's own classifier and, for CLIP ViT-B/32, with an ImageNet linear probe; r, Pearson correlation over the nine CLIP ViT-B/32 runs. (B) Distance to baseline: $1-$ Spearman ρ between the perturbed and baseline RDMs over 720 THINGS-fMRI concepts, at the onset, onset + 1, and final epochs (108 comparisons). RN50 uses re-estimated BatchNorm statistics (Appendix E).

Checkpoints. Each of the 36 perturbed runs is compared with its baseline at three epochs: the onset t, $t+1$, and the final epoch (89; CLIP ViT-B/32: 31). Together with the baselines at the same epochs, these are 136 checkpoints, each probed on 21 datasets (2,856 probes). Figure 4A adds ImageNet probes for cases T and K, including checkpoints saved within epoch 71 (Table 10), and Figure 9A measures CLIP ViT-B/32's accuracy with an ImageNet probe.

Datasets and splits. Table 6 lists the 21 datasets. Fourteen use the train and test splits of the VTAB task definitions (Zhai et al., 2020) and seven their official splits. Probes are fit on the training split and scored once on the test split; no test image is used for fitting or for selecting hyperparameters. Of our 29-dataset probe benchmark, the suite omits PCam, which was probed on only a subset of the checkpoints, and the seven ImageNet-based datasets (ImageNet, ImageNet-V2, -A, -O, -R, and -Sketch, and ObjectNet); ImageNet probes are reported separately. The five fine-grained datasets of Figure 4B are the transfer datasets whose classes are subordinate categories (Kornblith et al., 2019b).

Features and preprocessing. Probes are fit to frozen penultimate-layer features, extracted in evaluation mode without augmentation: for RN50, the 2,048-dimensional global average pool of the last stage, after BatchNorm re-estimation (Appendix E); for ViT-B/16, the 768-dimensional class token after the final layer normalization; for ConvNeXt-B, the 1,024-dimensional global average pool followed by the classifier head's layer normalization; for CLIP ViT-B/32, the 768-dimensional class token before the image–text projection. Images are converted to RGB and preprocessed with each model's own evaluation transform. For the three classifiers, the shorter side is resized to 256 pixels, a 224×224 center crop is taken, and ImageNet normalization is applied, with bilinear interpolation for RN50 and bicubic for ViT-B/16 and ConvNeXt-B. For CLIP ViT-B/32, the shorter side is resized to 224 pixels (bicubic), the center is cropped, and CLIP's normalization is applied. Features are not standardized. Dimensions that are constant over a dataset's training images are dropped. This happens only in RN50, whose pooled features are rectified: on Rendered SST-2 (rendered text) in every RN50 checkpoint, baselines included (up to 2,030 of the 2,048 units), and on most other datasets at every onset of label shuffle and at the early and mid onsets of image noise.

Probe. Each probe is a multinomial logistic regression (a softmax over all classes, also for the two-class Rendered SST-2) with an L2 penalty on the weights and an unpenalized bias. It minimizes the mean cross-entropy over the n training images plus $\|W\|^2/(2Cn)$, which is the objective of

scikit-learn's `LogisticRegression`. It is fit by full-batch L-BFGS (history 10, strong-Wolfe line search) in double precision on a GPU, starting from zero and stopping when no gradient entry exceeds 10^{-4} or after 20,000 iterations. The score is top-1 accuracy on the test split, and the transfer-probe mean is the unweighted mean of the 21 top-1 accuracies. All but 2 of the 2,856 fits converged (median 182 iterations). The two exceptions, CLIP ViT-B/32 with mid-onset uniform images at epoch 13 on VOC 2007 and on GTSRB, stopped at the cap, so their scores are lower bounds.

Hyperparameter selection. The one hyperparameter, the inverse regularization strength C, was selected independently for every probe, that is, for every combination of dataset and checkpoint, so a perturbed checkpoint and its baseline at the same epoch each have their own C. Following Kornblith et al. (2019b) and Radford et al. (2021), a class-stratified 20% of the training split is held out as a validation set, and probes are fit on the other 80% for $C \in \{10^{-6}, 10^{-4}, \ldots, 10^{6}\}$. Two refinement rounds then evaluate the midpoints in $\log C$ on either side of the best value so far, for a final resolution of a factor of $10^{0.5}$, and the C with the highest validation top-1 is kept (ties go to the smaller C). The probe is refit on the full training split at that C and scored on the test split. Fits during selection are warm-started from the neighboring C and capped at 200 L-BFGS iterations, because selection needs only a ranking of the candidates. The selected C spans the whole range from 10^{-6} to 10^{6} and differs between a perturbed checkpoint and its baseline in 75% of comparisons (94% at the onset, 68% one epoch later, and 63% at the final epoch), so a single C shared by all checkpoints would have regularized many perturbed probes differently from their baselines.

ImageNet probes. The 1000-class and 16-category probes of Figure 4A use the same recipe. They are fit on a fixed class-stratified subset of 100,000 ImageNet training images (100 per class, the same images for every checkpoint) and scored on the 50,000 validation images; the 16 categories are those of Figure 3. CLIP ViT-B/32's ImageNet probe (Figure 9A) also uses this recipe, fit on all 1,281,167 training images, except that its final fit starts from the selected validation-set solution rather than from zero.

Probe seeds. No probe has a seed to vary in the usual sense. Given a checkpoint's features, the fit is deterministic: the objective is convex, the optimizer is full-batch and starts from zero, and there is no shuffling, augmentation, or dropout. The only random choices are the 80/20 selection split and the 100,000-image ImageNet subset, each drawn once and shared by every checkpoint, so a perturbed probe and its baseline are fit, selected, and scored on identical images. Each probe was fit once. Re-fitting 11 probes from their stored features (five datasets on three checkpoints) selected the same C in all 11 and reproduced the test accuracy exactly in 9. The other two, both on KITTI distance, differed by 2 of its 711 test images (± 0.28 pp), because GPU floating-point reductions are not bitwise reproducible, so a fit stopped by a gradient tolerance can stop a few iterations earlier or later.

Uncertainty. The standard error of a top-1 accuracy p on n test images is $\sqrt{p(1-p)/n}$. The 95% intervals are ± 0.1 to ± 3.7 pp per dataset (widest for KITTI distance, with 711 test images), at most ± 0.30 pp on a transfer-probe mean, and at most ± 0.42 pp on the gap between a perturbed and a baseline mean. The gap interval adds the two variances, which is conservative when the two models' errors on the same test images are non-negatively correlated; per-image predictions were not stored, so the paired variance cannot be computed. By this interval, 43 of the 108 gaps in Figure 9B cannot be distinguished from zero: 27 of the 36 at the final epoch and 16 of the 36 one epoch after the onset.

Table 6: Datasets of the transfer suite (Figure 4). Train and test are the probe's fitting and scoring sets; the C-selection split takes 20% of Train (Appendix F). VTAB: the train and test splits of the VTAB task definitions (Zhai et al., 2020), whose validation split is not used; official: the dataset's own train and test splits. FGVC Aircraft is labelled at the variant level and probed on its train split, not train+val. PASCAL VOC 2007 has one image per annotated object box, cropped to the box. STL-10 uses its labelled images only. The ImageNet rows are the other probes: the 1000-class and 16-category probes of Figure 4A share one class-stratified 100,000-image subset of the ImageNet training set (100 per class); the CLIP ViT-B/32 ImageNet probe of Figure 9A uses the full training set.

Dataset	Task	Splits	Classes	Train	Test
Fine-grained					
Oxford-IIIT Pets	pet breeds	VTAB	37	2,944	3,669
Food-101	dishes	VTAB	101	68,175	25,250
Stanford Cars	car models	VTAB	196	8,144	8,041
FGVC Aircraft	aircraft variants	official	100	3,334	3,333
Flowers-102	flower species	VTAB	102	1,020	6,149
Other					
CIFAR-10	objects	VTAB	10	45,000	10,000
CIFAR-100	objects	VTAB	100	45,000	10,000
STL-10	objects	official	10	5,000	8,000
Caltech-101	objects	VTAB	102	2,754	6,084
PASCAL VOC 2007	objects (box crops)	official	20	7,844	14,976
DTD	textures	VTAB	47	1,880	1,880
SVHN	house numbers	VTAB	10	65,931	26,032
MNIST	digits	official	10	60,000	10,000
GTSRB	traffic signs	official	43	26,640	12,630
Country211	country of photo	official	211	31,650	21,100
Rendered SST-2	sentiment of rendered text	official	2	6,920	1,821
EuroSAT	satellite land use	VTAB	10	16,200	5,400
RESISC45	aerial scenes	VTAB	45	18,900	6,300
CLEVR count	number of objects	VTAB	8	63,000	15,000
CLEVR distance	distance to closest object	VTAB	6	63,000	15,000
KITTI distance	distance to closest vehicle	VTAB	4	6,347	711
ImageNet probes					
ImageNet, 1000-class	objects	subset	1000	100,000	50,000
ImageNet, 16-category	superordinate category	subset	16	100,000	50,000
ImageNet, CLIP only	objects	full	1000	1,281,167	50,000

Table 7: Transfer-probe mean (%, mean top-1 over the 21 datasets of Table 6) of the unperturbed baselines at every epoch that Figure 9 compares with a perturbed run. t is the onset epoch: 5, 30, and 70 for the ImageNet classifiers and 5, 13, and 28 for CLIP ViT-B/32; final is epoch 89 (CLIP ViT-B/32: 31).

	Early		Mid		Late		
Model	t	$t+1$	t	$t+1$	t	$t+1$	Final
RN50	65.6	66.1	68.5	68.3	69.8	69.8	70.5
ViT-B/16	47.3	50.3	66.9	66.8	72.3	71.7	72.1
ConvNeXt-B	64.1	65.7	71.3	71.0	72.9	72.8	73.0
CLIP ViT-B/32	54.1	56.8	67.0	67.6	70.3	70.5	70.4

Table 8: Transfer-probe mean of every perturbed run (%) and its gap to the baseline at the same epoch (perturbed − baseline, percentage points); the baselines are in Table 7. t, onset epoch; final, epoch 89 (CLIP ViT-B/32: 31). Test-set sampling alone puts at most ± 0.42 pp (95%) on any gap (Appendix F). Figure 9A plots the ratio of the two final-epoch means and Figure 9B the gaps.

			Perturbed			Gap		
Onset	Perturbation	t	t	$t+1$	Final	t	$t+1$	Final
RN50								
early	image noise	5	42.7	65.1	69.9	−22.9	−1.0	−0.6
early	label shuffle	5	22.8	65.5	70.2	−42.8	−0.6	−0.2
early	uniform images	5	56.6	65.6	70.4	−9.0	−0.5	−0.1
mid	image noise	30	49.3	68.1	70.5	−19.3	−0.3	0.0
mid	label shuffle	30	28.9	68.6	70.3	−39.6	+0.2	−0.1
mid	uniform images	30	55.9	68.0	70.1	−12.7	−0.3	−0.4
late	image noise	70	63.9	70.0	70.7	−5.9	+0.2	+0.2
late	label shuffle	70	37.9	70.0	70.4	−31.9	+0.2	−0.1
late	uniform images	70	65.7	69.9	70.5	−4.1	+0.1	0.0
ViT-B/16								
early	image noise	5	37.3	45.9	72.4	−10.0	−4.3	+0.3
early	label shuffle	5	36.5	45.5	72.5	−10.8	−4.8	+0.3
early	uniform images	5	37.0	46.2	72.4	−10.2	−4.0	+0.3
mid	image noise	30	51.6	66.9	72.5	−15.3	+0.1	+0.4
mid	label shuffle	30	36.9	66.8	72.5	−30.0	0.0	+0.4
mid	uniform images	30	43.2	66.3	72.6	−23.6	−0.5	+0.5
late	image noise	70	70.3	71.9	72.2	−2.0	+0.2	+0.1
late	label shuffle	70	50.3	71.9	72.3	−22.0	+0.2	+0.2
late	uniform images	70	70.9	72.0	72.4	−1.4	+0.3	+0.3
ConvNeXt-B								
early	image noise	5	32.7	64.1	72.7	−31.3	−1.6	−0.4
early	label shuffle	5	16.3	63.4	73.0	−47.8	−2.3	0.0
early	uniform images	5	37.0	63.7	72.6	−27.0	−1.9	−0.4
mid	image noise	30	11.1	18.2	72.8	−60.2	−52.9	−0.2
mid	label shuffle	30	11.1	70.3	72.7	−60.3	−0.8	−0.3
mid	uniform images	30	55.3	70.4	73.2	−16.0	−0.6	+0.2
late	image noise	70	53.9	73.0	73.3	−19.0	+0.2	+0.3
late	label shuffle	70	52.4	72.9	73.2	−20.5	+0.1	+0.2
late	uniform images	70	72.1	72.8	73.2	−0.8	0.0	+0.1
CLIP ViT-B/32								
early	image noise	5	38.3	19.2	64.1	−15.8	−37.5	−6.3
early	label shuffle	5	28.2	19.1	62.1	−25.9	−37.7	−8.3
early	uniform images	5	26.7	18.4	46.6	−27.5	−38.4	−23.8
mid	image noise	13	33.6	27.0	69.5	−33.3	−40.6	−1.0
mid	label shuffle	13	19.2	19.1	67.5	−47.8	−48.5	−3.0
mid	uniform images	13	24.5	20.9	66.7	−42.5	−46.7	−3.8
late	image noise	28	62.4	70.2	70.5	−7.9	−0.3	0.0
late	label shuffle	28	44.1	70.2	70.6	−26.2	−0.3	+0.1
late	uniform images	28	61.3	69.9	70.2	−9.0	−0.6	−0.3

Table 9: Per-dataset linear-probe top-1 (%) at the final epoch (89; CLIP ViT-B/32: 31) for the four baselines and for case L (CLIP ViT-B/32, image noise at epoch 5), and L's gap to its baseline (percentage points). ±, 95% test-set sampling interval of the gap (Appendix F). The gaps are Figure 4B.

	Baseline					
Dataset	RN50	ViT-B/16	ConvNeXt-B	CLIP	L	Gap of L
Fine-grained						
Oxford-IIIT Pets	92.2	92.7	93.2	73.0	55.5	-17.4 ± 2.2
Food-101	69.3	75.9	76.4	69.5	54.2	-15.3 ± 0.8
Stanford Cars	47.1	59.6	58.8	36.8	22.0	-14.8 ± 1.4
FGVC Aircraft	40.4	46.7	46.8	36.8	25.5	-11.3 ± 2.2
Flowers-102	80.3	86.8	80.1	87.7	76.7	-11.0 ± 1.3
Other						
CIFAR-10	87.6	94.3	93.9	92.3	89.7	-2.6 ± 0.8
CIFAR-100	69.2	79.5	78.4	76.9	72.0	-4.9 ± 1.2
STL-10	95.9	98.1	98.6	95.0	89.1	-5.9 ± 0.8
Caltech-101	88.5	88.9	88.1	88.9	81.7	-7.2 ± 1.3
PASCAL VOC 2007	81.6	84.1	84.6	82.4	79.2	-3.2 ± 0.9
DTD	68.3	68.8	69.8	67.9	53.6	-14.3 ± 3.1
SVHN	72.6	67.5	75.4	61.7	60.3	-1.4 ± 0.8
MNIST	98.3	98.3	98.5	98.0	97.8	-0.1 ± 0.4
GTSRB	78.5	76.7	79.9	81.0	75.2	-5.8 ± 1.0
Country211	12.1	12.4	11.7	15.0	11.2	-3.8 ± 0.6
Rendered SST-2	52.2	53.8	53.0	54.1	52.9	-1.3 ± 3.2
EuroSAT	96.3	96.6	95.6	97.3	96.0	-1.2 ± 0.7
RESISC45	88.5	89.4	88.0	89.4	84.3	-5.1 ± 1.2
CLEVR count	55.2	56.0	59.6	72.4	63.3	-9.1 ± 1.1
CLEVR distance	58.4	51.4	59.1	60.2	61.0	$+0.8 \pm 1.1$
KITTI distance	47.4	37.3	44.2	43.0	45.3	$+2.3 \pm 5.2$
Mean	70.5	72.1	73.0	70.4	64.1	-6.3 ± 0.4

Table 10: ImageNet linear-probe top-1 (%) of cases T (ViT-B/16, image noise at epoch 70) and K (ConvNeXt-B, label shuffle at epoch 70), perturbed / baseline, behind the 1000-class and 16-category readouts of Figure 4A. Checkpoints inside epoch 71 come from a re-run of that epoch saved after the listed number of optimizer steps (834 per epoch) and are compared with the baseline at the end of epoch 71; the epoch-71 row is the original run.

	T		K	
Checkpoint	1000-class	16-category	1000-class	16-category
end of epoch 70	67.2 / 75.3	82.7 / 85.4	0.1 / 80.2	66.3 / 90.5
epoch 71, step 11	74.0 / 75.4	85.0 / 85.3	0.1 / 80.3	69.6 / 90.5
epoch 71, step 19	74.3 / 75.4	84.9 / 85.3	42.8 / 80.3	75.8 / 90.5
epoch 71, step 36	74.5 / 75.4	84.9 / 85.3	77.4 / 80.3	88.7 / 90.5
epoch 71, step 67	74.5 / 75.4	84.7 / 85.3	79.5 / 80.3	89.6 / 90.5
epoch 71, step 126	74.6 / 75.4	85.0 / 85.3	79.7 / 80.3	90.0 / 90.5
epoch 71, step 236	74.7 / 75.4	84.9 / 85.3	79.8 / 80.3	90.0 / 90.5
epoch 71, step 443	74.8 / 75.4	85.0 / 85.3	79.9 / 80.3	90.1 / 90.5
epoch 71, step 834	75.0 / 75.4	85.2 / 85.3	80.0 / 80.3	90.3 / 90.5
epoch 71	75.0 / 75.4	85.3 / 85.3	79.9 / 80.3	90.3 / 90.5

G Robustness to the Alignment Metric

Every behavioral-alignment result in Section 4 holds when alignment is measured with CKA instead of RSA (Figure 10). Across the 36 perturbed runs, the two metrics agree on the retained fraction at the onset (Pearson $r = 0.93$) and on the gap to baseline at the final epoch ($r = 0.99$; supervised runs alone, 0.81). Under CKA, the same three late input-side runs exceed their baseline at the onset (ViT-B/16 with image noise, 1.38, and uniform images, 1.25; ConvNeXt-B with uniform

images, 1.04), and every supervised run ends within 0.014 of its baseline. Of the 24 supervised runs whose CKA fell by more than the baseline's epoch-to-epoch change, 23 regain half the drop within 0.4 epoch, sooner than their top-1 in 22 of those 23; ConvNeXt-B with mid image noise is again the exception. CLIP ViT-B/32 regains half its drop in the same number of epochs as under RSA, ends below baseline after all three early perturbations (down to -0.21 after uniform images), and ends at or above baseline after mid image noise and label shuffle. Category separability S predicts CKA retention as it predicts RSA retention (Figure 3D): $r = 0.59$, and partial $r = 0.49$ controlling for fidelity F and 0.50 controlling for 10-NN class purity, both with 95% confidence intervals excluding 0 ($[0.18, 0.71]$ and $[0.20, 0.71]$). As with RSA, the cross pair does not survive: fidelity's partial correlations with CKA retention are -0.05 and 0.10. CKA uses the same features as RSA (CLIP ViT-B/32: 768-d class-token features; RN50 with re-estimated BatchNorm statistics). In Figure 10D, runs whose alignment fell by less than the baselines' median change between consecutive epochs (0.01 for RSA, 0.0075 for CKA) are omitted, and ConvNeXt-B with mid image noise sits at its lower bound of 1 epoch under both metrics.

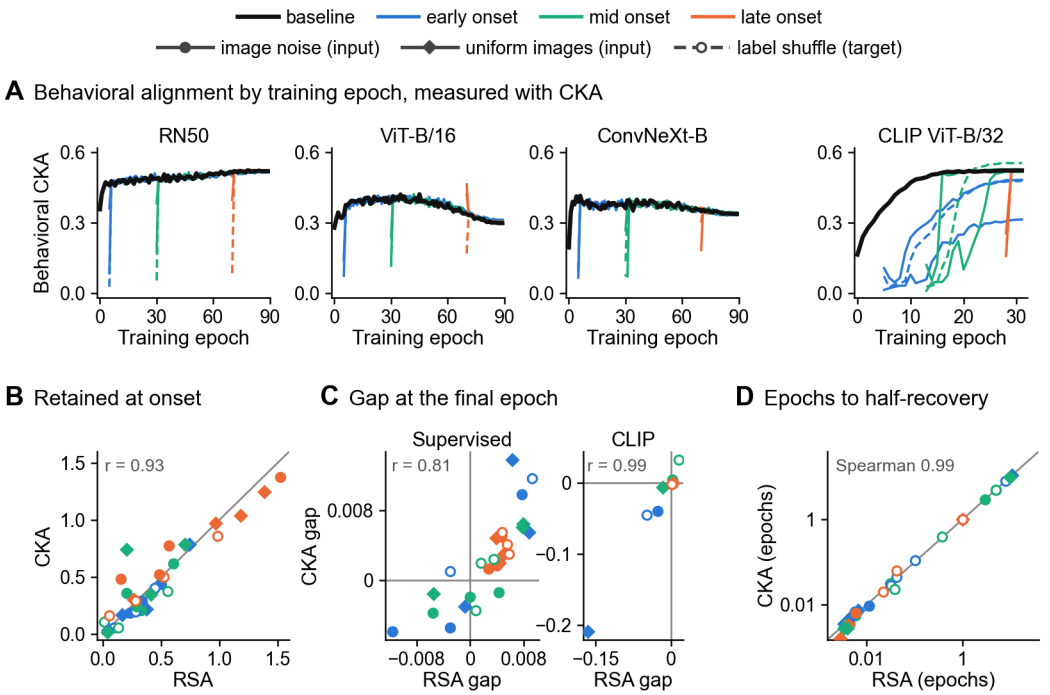

Figure 10: **Behavioral alignment measured with CKA reproduces the RSA results.** Linear CKA between penultimate-layer features and the SPoSE embeddings of the 1,854 THINGS images, at the same checkpoints and preprocessing as behavioral alignment. B–D compare the 36 perturbed runs under CKA and RSA; r, Pearson correlation. (B) Retained fraction at the onset; line, $y = x$. (C) Gap to baseline at the final epoch. (D) Half-recovery time as in Figure 6B, for runs timed under both metrics.

H Human Representational Alignment Trajectories at Baseline

Deep vision models often share properties with primate vision based on their architecture. Convolutional neural networks (CNNs), for example, mimic the hierarchical structure of the ventral visual stream. Early convolutional layers identify local, low-level features such as line orientation and edges before pooling this information in later layers to form more complex and holistic representations. Similarly, input to the primate ventral visual stream is first processed in V1 as highly localized receptive fields before downstream areas like V4 and the inferotemporal (IT) cortex integrate these features to construct higher-level representations. As a result, CNNs tend to align well with primate visual representations even early in training.

Vision Transformers (ViTs) (Dosovitskiy et al., 2021) have become foundational in modern computer vision, but they share fewer architectural similarities with the primate brain. Their multi-head attention mechanism means that any hierarchical information processing is learned, not imposed by architecture. As a result, human-ViT alignment emerges later in training, after seeing more data, and it relies more heavily on the training regimen applied. In addition, they have shown promising alignment with human and non-human primate vision. Dosovitskiy et al. (2021) showed that transformers can match or exceed convolutional neural network (CNN) performance on ImageNet given sufficient data, and subsequent work found that ViTs are more shape-biased and less texture-biased than CNNs (Naseer et al., 2021), similar to human inductive biases. The contrastive language-image objective that characterizes CLIP training (Radford et al., 2021) extended ViTs by training on 400 million image–text pairs with a contrastive objective, yielding representations that generalize broadly in a zero-shot setting. Additionally, Khanal et al. (2023) found that ViTs with large-scale pretraining are more resistant to label noise than CNNs trained from scratch, suggesting that pretrained representations can buffer against corruption. Such information can inform our reading of the baseline dynamics seen below.

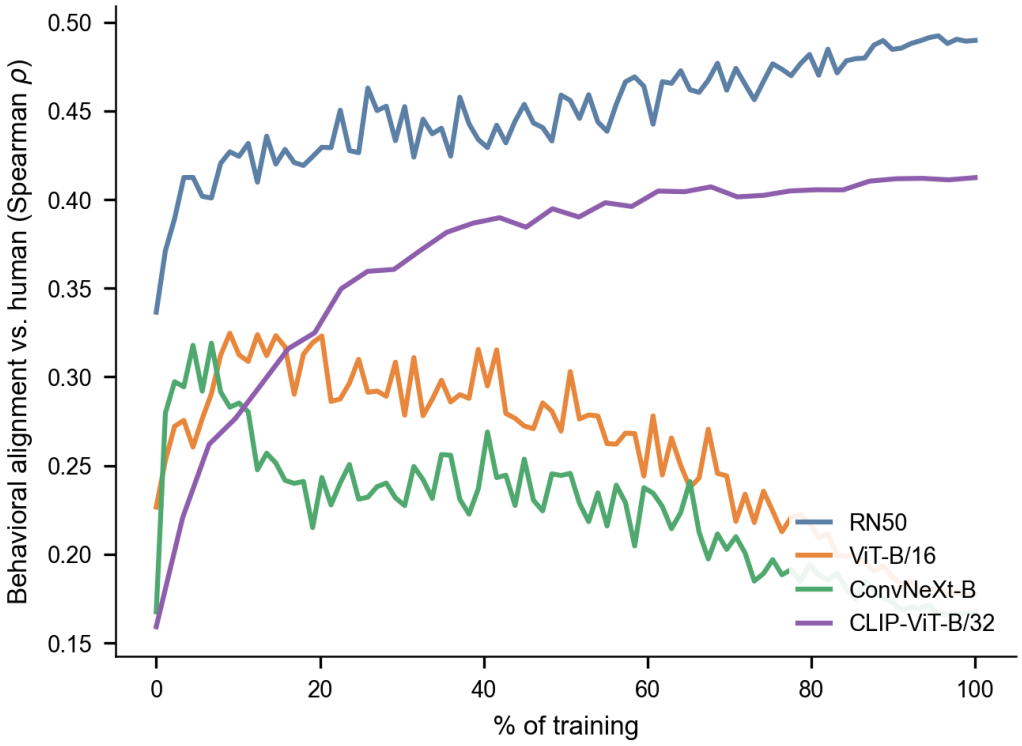

Figure 11: **Baseline alignment peaks early in ViT-B/16 and ConvNeXt-B but keeps rising in RN50 and CLIP ViT-B/32.** Alignment is measured using Spearman ρ against the 66-dimensional human SPoSE embeddings over the 1,854 THINGS images.

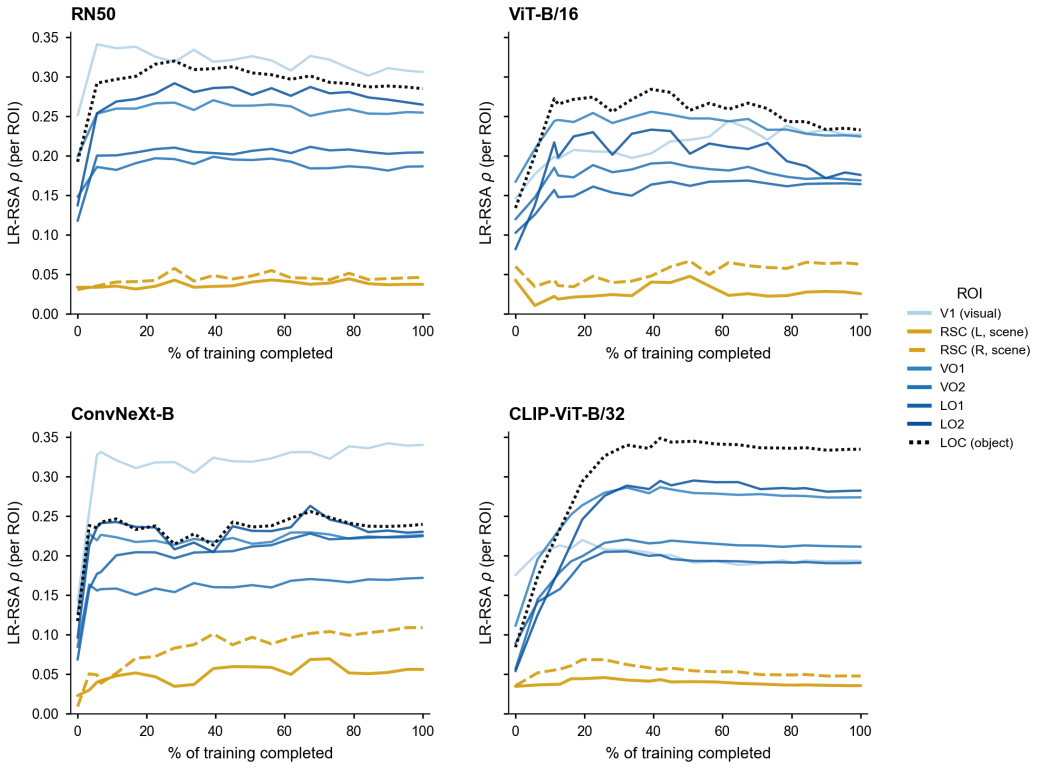

Figure 12: **On THINGS-fMRI, V1 aligns best with the convolutional baselines and LOC with the transformers.** Layer-reweighted RSA (Spearman ρ) between each baseline model at the given epoch and each ROI's subject-averaged reference RDM, for 8 of the 27 ROIs. Scores are in-sample, so absolute levels depend on each model's layer count; compare shapes within an ROI. Which region a model aligns to best splits cleanly by architecture family: at the final epoch V1 leads in both convolutional models, while object-selective LOC leads in both transformers. Scene-selective RSC sits below every other ROI in every architecture, so its resistance to perturbation should be read against a low starting level.

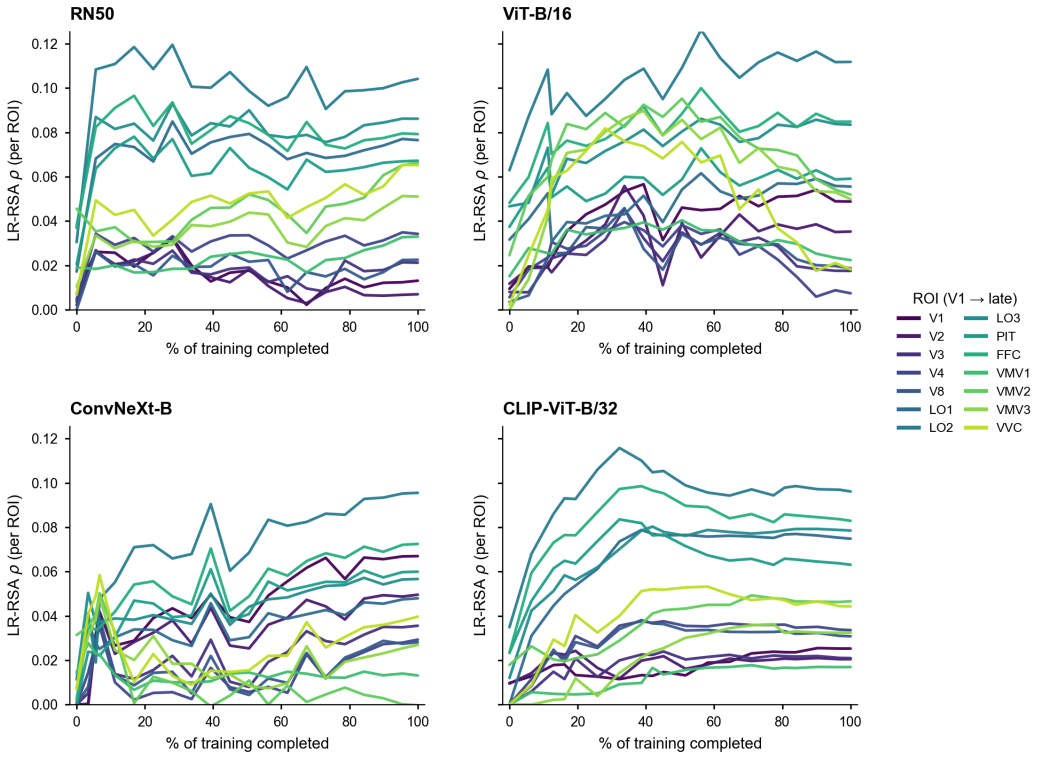

Figure 13: **On NOD, alignment is front-loaded and LO2 aligns best in every baseline.** Layer-reweighted RSA (Spearman ρ) between each baseline model at the given epoch and each of NOD's 14 subject-averaged ROI RDMs. Color runs V1 $\rightarrow$ high-level. Scores are in-sample, as in Figure 12. Alignment is front-loaded in all four architectures, and LO2 is the best-aligned ROI in each of them, both on average across training and at the final epoch.

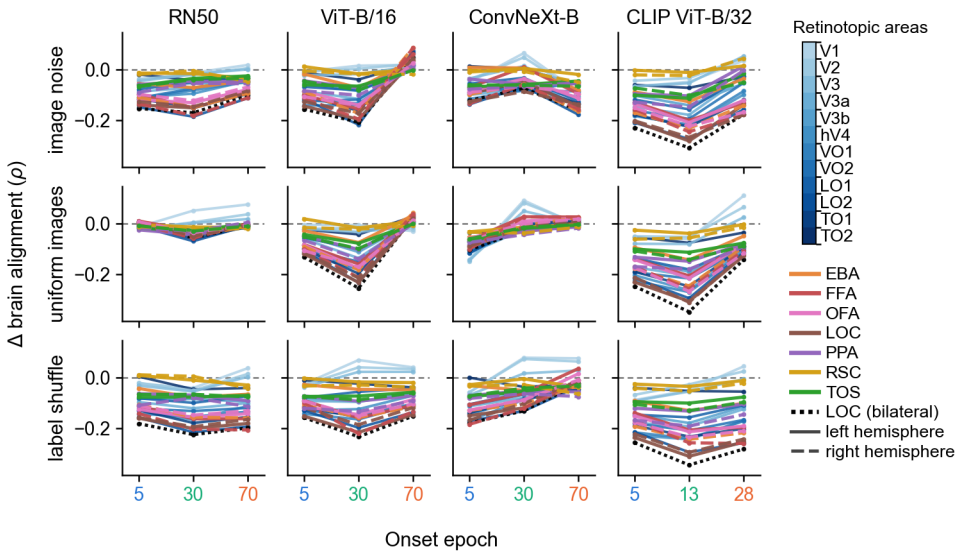

Figure 14: **Perturbations cost high-level visual cortex more alignment than early visual cortex.** Perturbed minus baseline layer-reweighted RSA (the per-ROI score that brain alignment averages; Section 3.2) at the end of the perturbed epoch, one line per ROI, with the baseline at the same epoch. Onsets as in Section 3.3. RN50 uses re-estimated BatchNorm statistics (Figure 2).

I Replicating Neural Alignment Results using the NOD Dataset

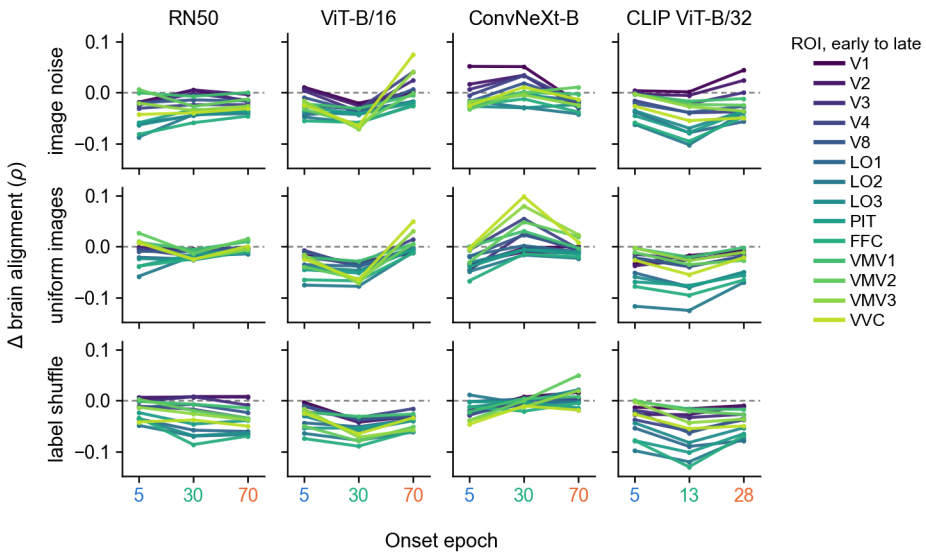

Figure 15: **The same hierarchical gradient replicates on NOD, an out-of-distribution stimulus set with an independent ROI parcellation.** Change in layer-reweighted RSA at the perturbed epoch relative to the same-architecture baseline, one line per ROI, ordered V1 $\rightarrow$ high-level along the colormap. The temporal ordering matches Figure 14 at roughly a third of the absolute magnitude, and the same hierarchical gradient holds: V1–V4 are close to unaffected while LO1–LO3, PIT, and FFC absorb most of the effect.

J Additional Supplementary Figures

In Figure 16, epoch 5 is the end of the perturbed epoch, and the frames match Figure 8A. The twelve pixel statistics are computed on the 224×224 pixels CLIP receives: luminance (mean of $Y = 0.299R + 0.587G + 0.114B$), RMS contrast (SD of Y), saturation (mean HSV saturation), colorfulness (Hasler & Süsstrunk, 2003), mean red, green, and blue, red–green balance (mean $R - G$), yellow–blue balance (mean $(R+G)/2-B$), edge density (mean Sobel gradient magnitude of Y), high-frequency power (share of Y's spectral power above half the Nyquist frequency), and entropy (of the 64-bin Y histogram). The CIFAR-10 probes (50,000 training and 10,000 test images) follow the transfer-probe protocol (Appendix F); the collapsed features are those of epoch 5; statistics and raw pixels ($32 \times 32 \times 3$) are z-scored, and CLIP features are used as is.

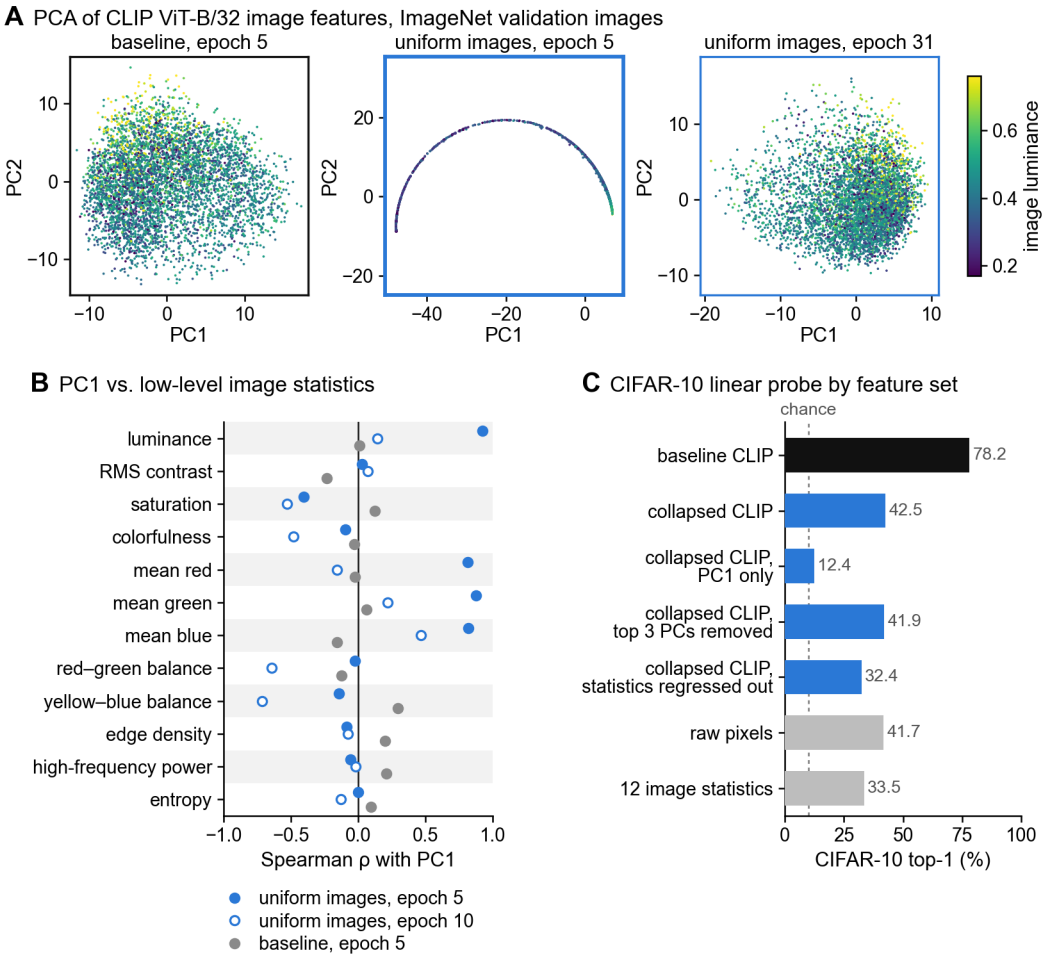

Figure 16: **One early epoch of uniform images collapses CLIP ViT-B/32's features onto a single luminance axis.** Uniform images at epoch 5 of 32; 768-d class-token features. (A) First two principal components of the features of the 50,000 ImageNet validation images (every 10th shown), fit separately for each model and colored by image luminance; PC1 is signed so that brighter images score higher. (B) Spearman correlation between PC1 and twelve pixel statistics (defined in Appendix J). (C) CIFAR-10 linear-probe top-1 on the baseline's features; on the collapsed features, whole, restricted to PC1, with the top three PCs projected out, or with the statistics regressed out; on the twelve statistics; and on raw pixels. Dashed line, chance.